VOLUME 4
TEST MANAGEMENT PHASE

CHAPTER 8
DESIGN OF EXPERIMENTS

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

19970117 180

SEPTEMBER 1994

USAF TEST PILOT SCHOOL
EDWARDS AIR FORCE BASE, CALIFORNIA

DTIC QUALITY INSPECTED 1

TABLE OF CONTENTS

8.1 GENERAL INTRODUCTION

The nature of testing lies in answering a question. This question may be phrased as

> "Does this system meet the specification requirements?"

or as

> "How much of what I taught was learned by the students?"

In both cases, a test is performed to determine the answer. This course material addresses the design of a test or experiment in order to satisfactorily answer the question posed to the tester. In the field of aerospace testing, some common questions posed to the test community include

> "We've developed a new system to meet the projected threat. Does this system meet the performance requirements in the specifications?"
>
> "We've modified this system and want to know whether the modification improves the system performance by a minimum measurable amount."
>
> "We have two systems that are being compared. Is one system measurably better than the other or are they substantially the same?"
>
> "Does the flight test data validate the analytical model derived from wind tunnel and analysis results?"

The tester is responsible for developing the test, from establishing the purpose of the test, establishing the success criteria, formulating the test hypothesis determining how much risk of being wrong is acceptable, and determining the minimum number of samples to take during the testing. Considering each of these factors in turn will enable the tester to ensure a successful test or experiment.

8.1.1 PURPOSE OF THE TEST

When presented with a problem that requires an answer to the test question, the tester can usually determine the purpose of the test. Whether it deals with a human factor evaluation of a new cockpit or a completely new aircraft, the test purpose should be stated at the start of planning. It's similar to stating the mission objective of an organization -- it provides the focus on a specific goal. Often, if the test purpose is not stated at the start of planning, the test scope and the success criteria become vague and generalized. The data collection can become confusing, since the purpose of collecting the data and the analysis of it has not been fully defined prior to the conduct of the test. Determining and stating the test's purpose usually provides a good map toward the successful accomplishment and conclusion of the test, and aids in the preparation of the test report.

Start test planning by
- first determining why the test is being performed, then
- stating "The purpose of this test is to...".

8.1.2 SUCCESS CRITERIA

Once the purpose of the test has been defined, the next step is to determine how you will know when the test is complete. The success criteria define this conclusion to the testing. Tests are conducted to gather data. These data may be as simple as preference between two types of hamburgers and may be collected visually or aurally, with nothing written down. In flight testing, however, the data collection system is often complex and expensive, collecting and recording data at phenomenal rates for subsequent analysis. The amount of various types of data to be collected can be the immediate product of testing and used to determine whether testing should be continued or not. These data can be construed as test success criteria.

Many times, success criteria begin with the words, "Collect sufficient data to...". It may seem odd to start with this success criterion, but consider the tester's intended objectivity.

The tester is not usually asked to show that one system is definitely superior to another or that a new system definitely meets the performance requirements. Instead, the tester's job is to subject the system(s) to test conditions in order to collect data. The analysis of these data should allow an objective conclusion regarding the original purpose of the test. It is a violation of a tester's ethical as well as legal responsibilities to bias the test setup, conduct, and analyses toward a predetermined conclusion. History is replete with examples of biased test setups, some resulting with a legal action against the tester.

Hence, the truly objective success criterion is merely to collect sufficient data in order to evaluate the article/process/person being tested. Once those data are obtained, an objective analysis will provide a repeatable audit trail from test definition through reporting the conclusions in the report. Repeatability of the test and objectivity of the tester are critical components of a valid test program.

If the success criteria reads, "Collect sufficient data to evaluate whether the ______ system meets the minimum specification requirements contained in XXXXXX.", how does the tester determine which data and when "sufficient data" have been collected?

That's one of the purposes of this course and will be covered under the determination of sample size. First, there are several more factors that should be considered; these are the test hypotheses and the amount of acceptable risk for being wrong.

8.1.3 TEST HYPOTHESES

A hypothesis is merely a statement that may or may not be true. A statistical hypothesis is a statement that a tester or experimenter makes and then gathers data to determine whether the data either supports the hypothesis or not. For the purposes of this course, we will consider two types of hypotheses: the null hypothesis and the alternate hypothesis.

The null hypothesis can sometimes be formulated from the purpose of the test. For example, if the tester is asked to determine whether a new aircraft meets the roll rate requirement at a particular altitude/airspeed/control deflection condition, the purpose of the test could be stated that way. The null hypothesis could be written as

H_0: the aircraft roll rate is at least xxx degrees per second at that condition.

For the purposes of this course, H_0 will designate the null hypothesis. The tester would then formulate the alternate hypothesis and then collect data at that condition and then use the analysis tools from this course to determine whether or not the data support the null hypothesis.

The alternate hypothesis is formulated to account for the possibilities not included in the null hypothesis. In the above example, the alternate hypothesis could be described as

H_1: the aircraft roll rate is less than xxx degrees per second at that condition.

where H_1 is the designation for the alternate hypothesis. Hypothesis formulation and testing is the subject in a subsequent section of this course and will be covered in detail there. The purpose of bringing to your attention here is to demonstrate how the formulation of test hypotheses is integrated into test development.

8.1.4 ACCEPTABLE RISK

Once the test hypotheses are formulated, the data collected and analyzed, and conclusions are drawn, the subject of being wrong must be addressed. Tests collect data in samples from populations. It would be impractical in most cases and impossible in some to measure each member of a population for the performance or characteristic of interest. Instead, samples are taken from the population and conclusions are drawn based on the data from those samples. There are risks incurred from using sample data and extrapolating to the population.

The two types of risks or errors that can occur are:

TYPE I: Reject the null hypothesis when it is in fact true (occurs with probability α; $\alpha \equiv$ level of significance).

TYPE II: Accept the null hypothesis when it is in fact false (occurs with probability β; $1-\beta \equiv$ power of test).

Note that both errors cannot occur simultaneously. The tester must determine how much risk he/she can accept for both types of errors in order to determine how many samples are sufficient to evaluate the system.

This course addresses the statistical and probability tools that can be used to enable the student to make pretest determinations and assumptions regarding sample size and to draw conclusions using quantitative assessments from sample data.

8.1.5 SAMPLE SIZE DETERMINATION

With an unlimited budget and no schedule constraints, testing can go on indefinitely. However, the reality is a test program that is always cost and schedule constrained, and, therefore an assessment of the required sample size is necessary. If insufficient samples are taken, then you may not be able to achieve your required confidence level. If too many samples are taken, then resources may be wasted.

Many senior officers have taken acquisition courses that have included sample size determination methods such as

- Latin square
- Full Factorial
- Partial or fractional factorial
- Taguchi method

The particular application of these techniques is test dependent

8.2 MEASUREMENT INACCURACIES

Flight testing consists almost entirely of experimental observations from which we record numbers: time to climb, fuel flow, short period frequency, Cooper-Harper ratings, and INS drift rate to name a few. All experimental observations have inaccuracies. Understanding the extent of these errors and developing methods to reduce their magnitudes to acceptable levels are the subjects of this course.

8.2.1 TYPE OF ERRORS

In discussing the errors in our experimental observations, we need to make a distinction between two very different kinds of errors: systemic errors and random errors.

Systemic errors are repeatable errors caused by some flaw in our measuring system. For example, if we measure lengths with a ruler that has the first inch broken off, our data will all have a one inch systemic error. The instrument corrections we apply to indicated airspeed and altitude to obtain calibrated airspeed and altitude are examples of compensating for known systemic errors. Systemic errors are often referred to as bias errors.

Random errors are not repeatable. If we make multiple observations of the same parameter with the same equipment under the same conditions, we will still have small variations in the results. These variations are caused by unobserved and uncontrollable changes in the experimental situation. They can result from small errors in the judgment of the observer, such as in interpolating between the marks of the smallest scale division of an instrument. Other error sources could be unpredictable variations in temperature, voltage, or friction. Because these errors are not repeatable, they can never be eliminated. Empirically, however, it has been found that such random errors are frequently distributed according to a simple law. Therefore, it is possible to use statistical methods to deal with these random errors to obtain meaningful results. Random errors are often referred to as precision errors.

8.2.2 TYPES OF DATA

All data are not of the same type. When we use a scale of one to ten to quantitatively rate the handling qualities of an aircraft, these data cannot be mathematically treated in the same way that we treat miss distance data on the bombing range. In fact there are four different types of data: nominal, ordinal, interval, and ratio data.

Nominal data are labels that are numerical in name only. Federal stock numbers for different supplies are an example. We cannot treat these data with any of the normal arithmetic processes. For instance, we cannot say that $3 > 1$ or that $3 - 1 = 2$ or that $4 \div 2 = 2$. With nominal data, none of these arithmetic operations are applicable.

Ordinal data contains information about rank order only. If we rank order different aircraft by their maximum speed, then the resulting data can be compared to say that for example, 3 > 1 meaning aircraft three is faster than aircraft one. We cannot, however, say that 3 - 2 = 1, or that 4 - 2 = 2. Ordinal data can only be used to reflect inequalities between the data.

Interval data contains the rank order information of ordinal data plus difference information. For example, temperature data has rank and difference information. If it is 30°F, 45°F, and 60°F at different times, the successive differences in temperature are the same, 15°F. In both cases, the same amount of heat had to be added to raise the temperature by 15°F. We cannot say, however, that the end temperature of 60° is twice as hot as 30°. The reason is that our zero point is arbitrary. Zero degrees Fahrenheit does not mean the absence of temperature. Thus, interval data has relative and difference information, but not ratio information.

Ratio data contains the information necessary to perform all the basic mathematic operations on the data. Most of our data falls into this category. Airspeed, fuel flow, range, etc, data all can be compared relatively, subtracted, and divided. We can legitimately say that a 1000 NM range in an F-4 is 4 times as great as a 250 NM range in an A-37.

This distinction between nominal, ordinal, interval, and ratio data is important. The type of data we have in a particular case may dictate the use of certain statistical techniques. But, before we can develop and use these statistical methods, we must first establish a common base of understanding of elementary probability.

8.2.3 ABBREVIATIONS AND SYMBOLS

The following abbreviations and symbols will be used in this text:

H_0	null hypothesis
H_1	alternate hypothesis
n	number of samples
P(A)	probability of event A occurring
$P(\bar{A})$	probability of event A not occurring
s	sample standard deviation

U	rank sum statistic
W	sign rank statistic
$\bar{x}$	sample mean
$\tilde{x}$	sample median
$\hat{x}$	sample mode
z	standard normal deviate
α	probability of type I error
β	probability of type II error
δ	difference in means
μ	population mean
ν	degrees of freedom
σ	population standard deviation

8.3 ELEMENTARY PROBABILITY

A quantitative analysis of the random errors of measurement in flight testing (or any other experiment) must rely on probability theory. Probability theory is a mathematical structure which has evolved for the purpose of providing a model for chance happenings. The probability of an event is taken to mean the likelihood of that event happening. Mathematically, the probability of event A occurring is the fraction of the total times that we expect A to occur, or

$$P(A) = \frac{n_a}{N}$$

Where: P(A) is the probability of A occurring n_a is the number of times we expect A to occur.

N is the total number of attempts or trials

From this definition, it can be seen that P(A) will lie between zero and one, since the least that n_A can be is zero (A never happens) and the most it can be is N (A always happens).

In order to determine this fraction, n_A/N, we can approach the problem in two distinctly different ways. We can use our previous knowledge and make assumptions to predict the probability - classical or 'a priori' probability - or we can conduct experiments to determine the probability - experimental or 'a posteriori' probability.

8.3.1 CLASSICAL PROBABILITY

The study of classical probability began hundreds of years ago when games of chance became fashionable. There was much interest in questions about how frequently a certain type of card would be drawn or that a die would fall in a certain way. For example, it is almost obvious that if an ideal die (six sided) is honestly cast, there are six possible outcomes and the chance of getting a particular face number is one out of six; i.e., the probability is 0.16667.

The underlying conditions for simple evaluations like this are that:

1. every single trial must lead to one of a finite number of known possible outcomes, and
2. each possible outcome must be equally likely.

If we satisfy these two conditions, then the probability of event A is just

$$P(A) = \frac{n_A}{N}$$

where now: n_A is the number of ways A can happen.

N is the total number of possible outcomes.

For example, what is the probability of getting no heads when we toss two fair coins? The possible outcomes are:

(H,H) (H,T) (T,H) (T,T)

Thus, N = 4 (that is four distinct, equally likely results) and $n_A = 1$ (only the result T,T has no heads).

Therefore, P(no heads) = 1/4 = .25

This approach to determining probabilities is instructive, but, in general, it is not applicable to experimental situations where the number of possible events is usually infinite and each possible outcome is not equally likely. Thus, we turn to experimental ('a posteriori') probability.

8.3.2 EXPERIMENTAL PROBABILITY

By definition, experimental probability is

$$P(A) = \lim_{N_{obs} \to \infty} \frac{n_{A_{obs}}}{N_{obs}}$$

where now: $n_{A_{obs}}$ is the number of times we observe A.

N_{obs} is the number of trials

For example, suppose we wish to check the classical result that the probability of getting a head when tossing a coin is 1/2. We toss the coin a large number of times and keep a record of the results. A typical graph of the results of such an experiment is shown in Figure 8.1. We will never, of course, reach an infinite number of trials, but our confidence in the probability of getting a head will increase as the number of trials increases. As can be seen in Figure 8.1, the fraction of observed heads fluctuates dramatically when N is small, but as N increases, the probability steadies down to an apparently equilibrium value.

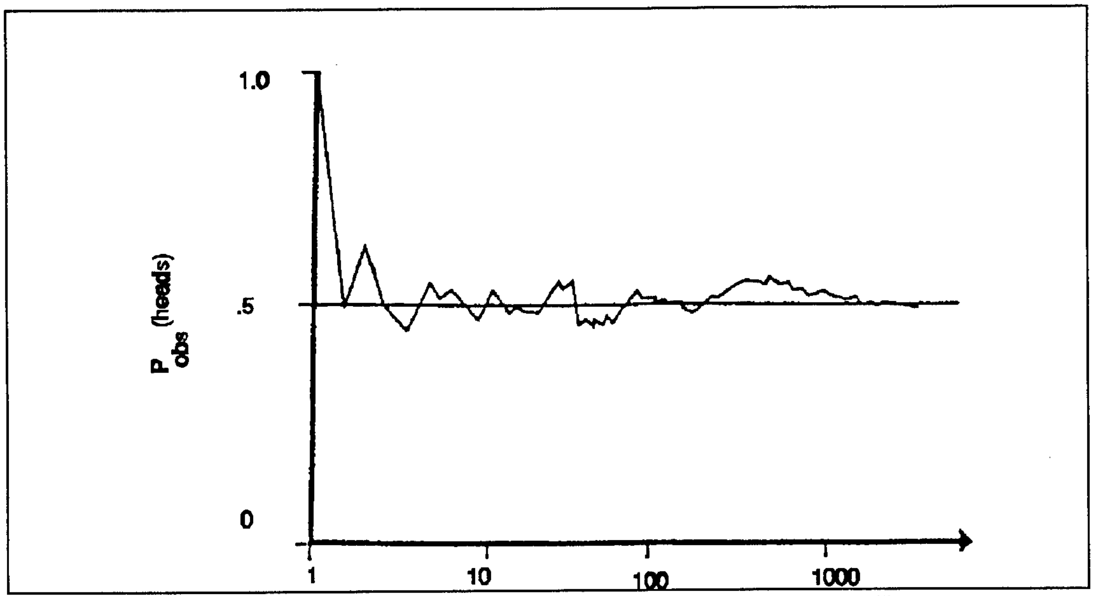


FIGURE 8.1 EXPERIMENTAL PROBABILITY

8.3.3 PROBABILITY AXIOMS

Probability theory can be used to describe the relationships between multiple events. Several axioms are presented. First, if the probability of A occurring is P(A), then the probability of A not occurring, $P(\overline{A})$, is just:

$$P(\overline{A}) = 1 - P(A)$$

This is easy to accept without a rigorous proof since the sum of the probability of something occurring and not occurring has to be one.

The remaining axioms presented below for multiple outcomes assume that each outcome is independent (A occurring does not subsequently affect the probability of A or B occurring) and mutually exclusive (only one can occur in a single trial). The two remaining axioms are:

$$P(\text{A or B}) = P(A) + P(B)$$

$$P(\text{A and B}) = P(A) \times P(B)$$

These axioms are also easily justified (as opposed to proven) by looking at classical probability. If we take the example of tossing a coin, then

$$P(H) = .5 \qquad P(T) = .5$$

and

$$P(H \text{ or } T) = .5 + .5$$

which makes sense, because the probability of the coin coming up either heads or tails has to be one (excluding the chance of landing on edge).

Also, from the example of getting two tails in section 8.3.1.

$$P(T \text{ and } T) = P(T) \text{ x } P(T) = .5 \text{ x } .5 = .25$$

which is the same answer we got by examining all of the possible outcomes.

8.3.4 PROBABILITY EXAMPLES

Problem: Based on historical data, suppose we determine that 95% of the time an F-4 will make a successful approach end barrier engagement. If we have a flight of four that must use the barrier due to icy runway conditions, what is the probability that at least one aircraft will miss the barrier?

Solution: The probability that at least one will miss is the complement of the probability that all will successfully engage. That is

$$P(1 \text{ or more miss}) = 1 - P(\text{all engage})$$

The probability that all four engage is

$$\begin{aligned} P(\text{all engage}) &= P(\text{1st engages}) \text{ x } P(\text{2nd engages}) \text{ x } P(\ldots.. \\ &= P(\text{single engagement})^4 \end{aligned}$$

Finally, since P(each engages) = .95,

$$P(1 \text{ or more miss}) = 1 - (.95)^4 = 1 - .81 = .19$$

Or roughly speaking, about one out of five times, a flight of four F-4s would have at least one barrier miss.

Problem: What is the probability of getting craps (rolling a total of 2, 3, or 12) on a single roll of a pair of dice?

Solution: Since getting 2, 3 or 12 are independent, mutually exclusive events, we can use the following:

P(2, 3, or 12) = P(2) + P(3) + P(12)

To get individual probabilities, first note that there are 36 possible outcomes (6^2) with two dice, each having six sides. In order to get a total of 2 or 12, there is only one way the dice can come up: 1 and 1 or 6 and 6, respectively. For a total of 3, the dice can come up two ways: 1 and 2 or 2 and 1. Therefore, since $P(A) = n_A/N$, we have:

P(2) = 1/36
P(3) = 2/36
P(12) = 1/36

and finally

P(2,3, or 12) = 1/36 + 2/36 + 1/36 = 1/9

Thus, about 11% of the time that you roll the dice, you will crap out.

8.4 POPULATIONS AND SAMPLES

8.4.1 DEFINITION

Thus far in our discussion, we have made no distinction between populations and samples. The difference is an important one in the study of statistics. The definitions follow.

A data population is all possible observations of a certain phenomenon. Thus, many populations are infinite. For example, the population of the totals of two dice are all possible (infinite) outcomes of rolling two dice. Another example, the population of weapon deliveries from an aircraft is all the possible drops it could make in its lifetime. A more limited population would be the scores of your class on the final exam. This population would have only limited possible observations, not infinity. Characteristics of a population, such as its mean, variance, or mode, are called parameters.

A data sample is any subset of a given population. Thus the results of 1000 rolls of two dice constitute a sample of all possible results. The scores from 100 F-4 sorties could be another example. The scores of 5 of your classmates would be a sample of the results of the whole class. Characteristics of a sample are called statistics.

8.4.2 ASSUMPTIONS

Constructing a population (what should be included as possibilities, what should be excluded?) or selecting a sample from a population must be done with care if we are later to apply statistical analysis techniques. The assumptions we normally impose on samples are that the data be homogeneous, independent, and random.

A homogeneous sample has data from one population only. If, for example, we allow bomb scores from an F-4C (iron sight) and an F-16C (predictive heads-up display) to be included in a single sample, the results would not be very meaningful.

An independent sample is one where the selection of one data point does not affect the likelihood of subsequent data points. For example, after dropping a bomb thirty feet long on the first pass, the probability that the next drop will miss by the same distance (or any other distance) is unaffected by the last sample (independent). An example where the subsequent probabilities are affected is sampling from a finite population without replacement. For example, the probability of drawing a heart from a deck of cards changes if you sample and discard. The sample would remain independent if you replaced the card after each draw.

A random sample is one where there is an equal probability of selecting any member of the population. An example of a nonrandom sample would be using a single F-16 with a boresight error causing a bias in downrange miss distance to produce samples intended to be representative of all F-16 weapon deliveries.

8.4.3 MEASURES OF CENTRAL TENDENCY

Given a homogeneous, independent, random sample, we now turn to methods to describe the contents of that sample. Suppose, for instance, we wish to be very accurate in measuring a hard steel rod with a micrometer. The population of measurements is all of the possible measurements that could be made with the micrometer. If we take a sample of ten measurements, we will probably get several different answers. The unpredictable variations could come from any of several different sources: we may tighten the micrometer sometimes more than others, there may be small dust particles sometimes, we may make small errors in estimating tenths of the smallest scale division, and so forth. Even so, we would expect to get a better answer by measuring many times rather than just once.

But what should we do with the multiple measurements, some of which are different? The most obvious procedure would be to average them. When we average the contents of a sample, we call the result the arithmetic mean, usually denoted by $\overline{x}$:

$$\bar{x} = \frac{1}{N} \Sigma x_i$$

The mean is the most common measure of central tendency, but not the only one. If we had taken 10 measurements and 8 of them were the same, we might feel justified in stating that this most common answer is the correct one and that the other 2 different answers were due to some unseen error. Using the most common sample is called taking the mode. The mode (usually denoted $\hat{x}$) is the most frequent sampled value. In some samples, there may be more than one mode.

A third measure of central tendency is the median. The median (usually denoted $\tilde{x}$) value is the middle value. If we rank order the sample elements, then for an odd number of elements the median is just the middle value. For an even number of elements, we define the median as the arithmetic average of the two middle values.

Of the three different measures of central tendency (mean, mode, and median), the mean or arithmetic average is most commonly used.

8.4.4 DISPERSION

Given that we usually use the mean, $\bar{x}$, as the single measure of central tendency of a sample, is that enough to adequately characterize the contents of a given sample? The answer is no. Using the mean by itself can be very misleading. For instance, consider the following two samples:

Sample 1: 99.9, 100, 100.1

Sample 2: 0.1, 100, 199.9

As can be seen, the mean (and median in this case) is the same ($\bar{x} = \tilde{x} = 100$) for both samples yet there is a significant difference between these two samples. The difference is in the variation or dispersion of sample elements from the mean. Thus, we now need some measure of the dispersion within a sample.

To obtain a measure of dispersion, first define the deviation, d_i, as the difference between the ith element of the sample and the sample mean:

$$d_i = x_i - \overline{x}$$

The first inclination may be to average these deviations, but the result is not illuminating, since:

$$\overline{d} = \frac{1}{N} \sum_{i=1}^{N} d_i = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i - \overline{x}$$

$$= \overline{x} - \overline{x} = 0$$

Because of the definition of the mean, the deviations above the mean will always exactly cancel the deviations below the mean. This result may lead you to conclude that we should average the absolute values of the individual deviations. Doing so produces what is referred to as the mean deviation:

$$\textit{mean deviation} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|$$

This quantity is sometimes used as a measure of dispersion, but for reasons that will become apparent later on, a more common measure of dispersion is the standard deviation, which is defined next.

In defining the standard deviation, we eliminate the negative individual deviations by squaring each term, rather than by taking the absolute values. We then average the squares and finally take the positive square root of the results. Thus, the standard deviation (denoted by σ) is the root-mean-square deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (d_i)^2}$$

The square of the standard deviation, σ^2, is called the variance. The measures of central tendency described above are called descriptive statistics. They describe what's going on. Inferential statistics, which we'll describe later, deal with projecting sample statistics to characterize population parameters.

8.4.5 NOTATION

Normally, we use Greek letters to denote statistics (such as mean and variance) for populations and we use Roman letters for statistics of samples.

Therefore, we will use:

μ and σ^2 for population mean and variance

$\bar{x}$ and s^2 for sample mean and variance

There is one other difference between population and sample statistics. The sample standard deviation is defined slightly differently than the population standard deviation:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

The difference is that the sum of the squares is divided by N - 1 for the sample rather than by just N as for the population standard deviation. The effect is to make the sample standard deviation slightly larger, and more conservative, than it would have been and the difference decreases as the sample gets larger.

8.4.6 EXAMPLE

Problem: Given the following 10 observations find the sample mean, median, mode, and standard deviation: 3, 4, 6, 6, 6, 8, 9, 10, 12, 15

Solution:

$$\bar{x} = \frac{1}{10}(3 + 4 + 6 + 6 + 6 + 8 + 9 + 10 + 12 + 15) = 7.9$$

$$\hat{x} = 6 \ (\textit{most frequent})$$

$$\tilde{x} = \frac{1}{2}(6 + 8) = 7 \ (\textit{average of two middle values})$$

$$s = \sqrt{\frac{1}{9}(4.9^2 + 3.9^2 + 1.9^2 + 1.9^2 + 1.9^2 + .1^2 + 1.1^2 + 2.1^2 + 4.1^2 + 7.1^2)}$$

$$= 3.695$$

8.5 PROBABILITY DISTRIBUTIONS

Now that we have covered elementary probability concepts and introduced the idea of population and samples, we turn to probability distributions. Application of statistical methods requires an understanding of the characteristics of the data obtained. Probability distributions, either empirical or theoretical can give us these required characteristics. Most statistical methods are based on theoretical distributions which approximate the actual distributions.

8.5.1 DISCRETE PROBABILITY DISTRIBUTIONS

To introduce the idea of a probability distribution, let's go back to the example of tossing two coins in Section 13.2.1. We can calculate from classical methods the probability of getting 0, 1, or 2 heads. Tabulating this as f(n), where n is the number of heads:

Table 1. Probability of getting n heads in two tosses of a fair coin.

n	f(n)
0	.25
1	.50
2	.25

Another method of presenting this data would be graphically by means of a bar graph, as shown in Figure 8.2.

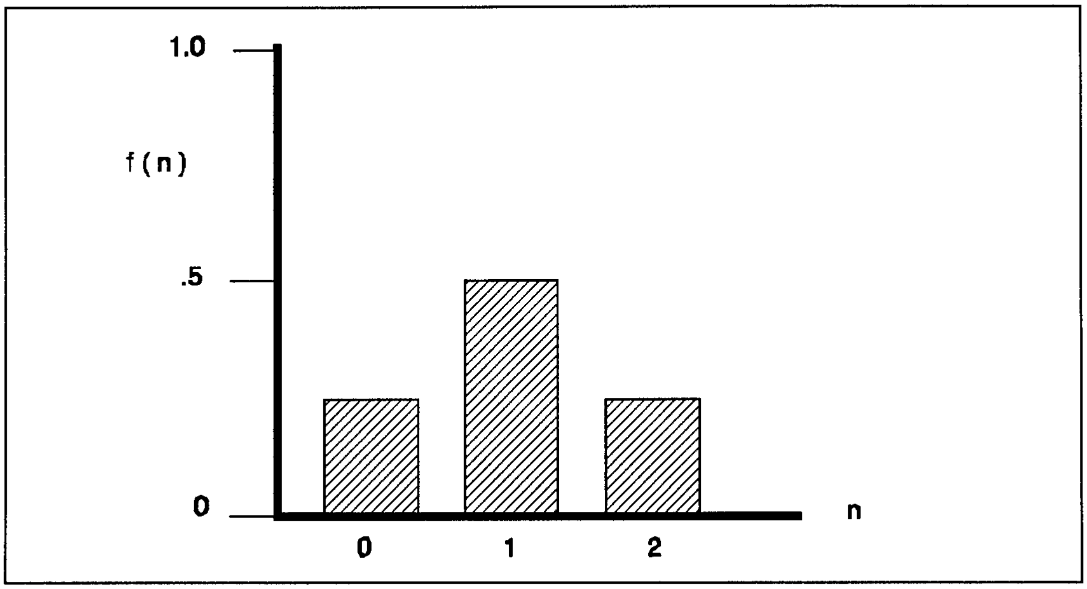


FIGURE 8.2 PROBABILITY OF GETTING N HEADS IN TWO TOSSES OF A FAIR COIN.

Thus, f(n) is called the probability distribution of n. The above example is a theoretical calculation. More frequently, we are concerned with empirical distributions. For example, suppose we collect a sample of data on T-38 landings as shown in Table 1.

TABLE 1. TOUCHDOWN DATA

Touchdown Distance from Aim Point	Frequency in Distance Interval	Relative Frequency
0 - 100 ft	2	.05
101 - 200 ft	10	.24
201 - 300 ft	18	.44
301 - 400 ft	8	.20
401 - 500 ft	3	.07

Plotting the data in a histogram as in Figure 8.3 will give us a graphical representation of this empirical probability distribution.

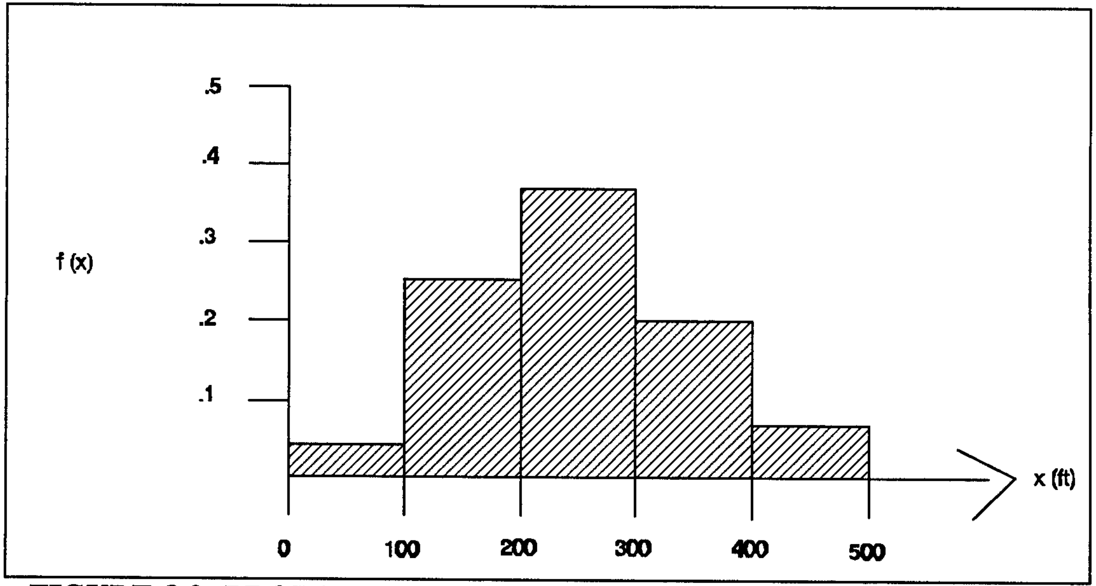


FIGURE 8.3 PROBABILITY DISTRIBUTION OF TOUCHDOWN MISS DISTANCE

8.5.2 CONTINUOUS PROBABILITY DISTRIBUTION

If we acquire more data on T-38 landings and reduce the size of the intervals, we could draw a new histogram. In the limit as we acquire more and more data, and reduce the interval size to smaller and smaller values, the histogram approaches a smooth curve, as shown in Figure 8.4.

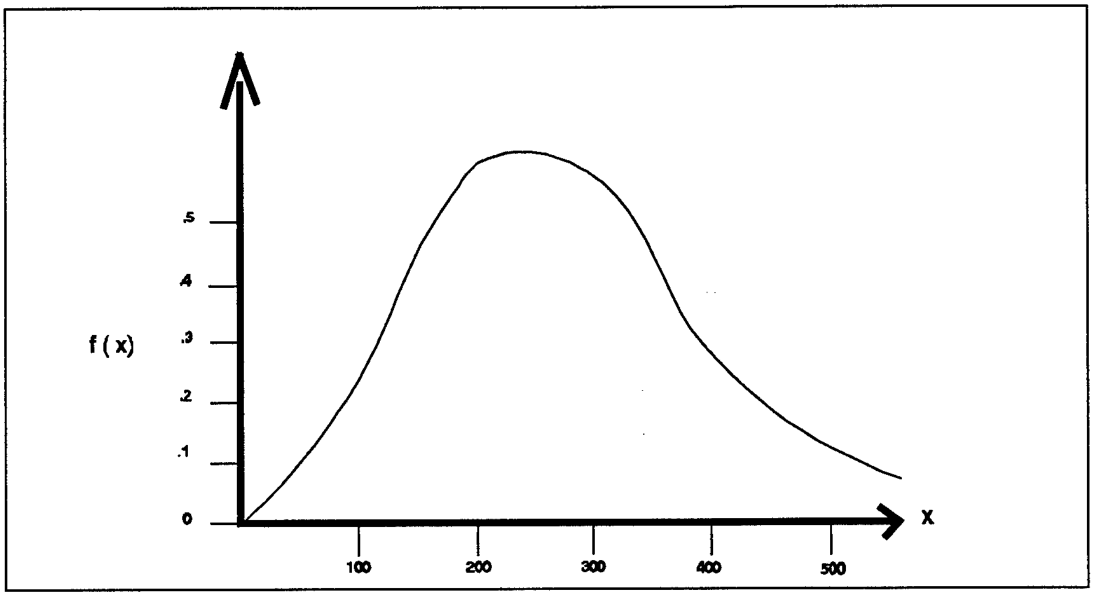


FIGURE 8.4 CONTINUOUS PROBABILITY DISTRIBUTION OF TOUCHDOWN MISS DISTANCE.

This smooth, continuous probability distribution cannot be interpreted in the same way as the discrete distribution. In Figure 8.3, the height of the bar above the interval is the probability that x will have a value within that interval. In Figure 8.4, the height of the curve above a point is not the probability of x having that point value. Since there are an infinite number of points (i.e., a continuous curve) the probability of x having a single specific value is zero. We can, however, talk about the probability of x being between two points, a and b. Then, the interpretation of the continuous probability distribution is as follows:

$$P(a \leq x \leq b) = \int_a^b f(x)\ dx$$

That is, the probability that x falls between a and b is the area under the probability distribution curve between x = a and x = b, as shown in Figure 8.5.

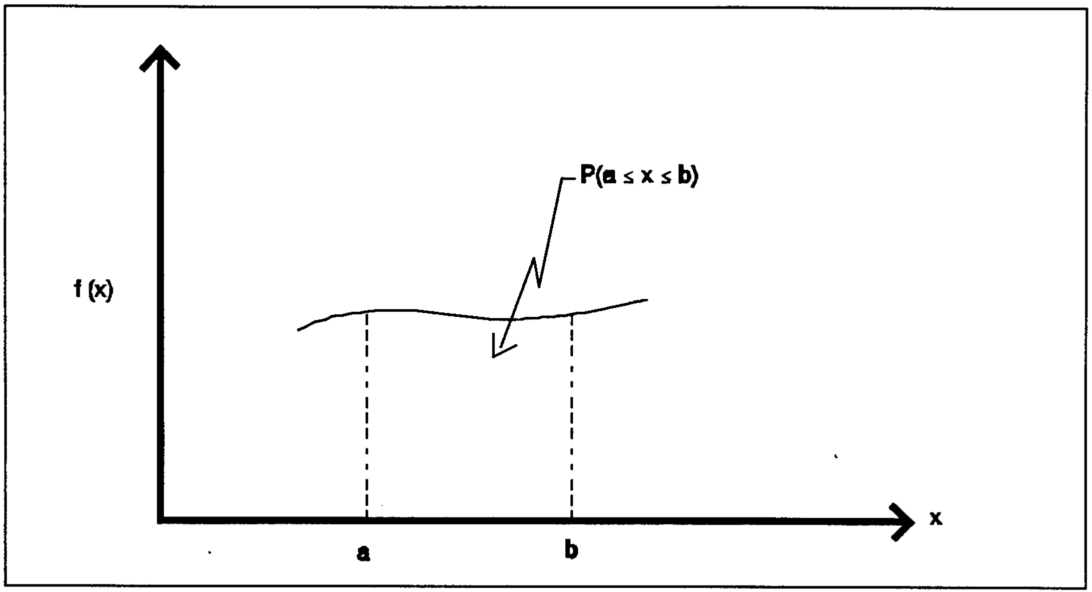


FIGURE 8.5 PROBABILITY VS. THE AREA UNDER A CONTINUOUS PROBABILITY DISTRIBUTION.

From this, we can see that f(x) must always be greater than or equal to zero. Negative areas would be meaningless. Also, since the maximum probability is one, we have:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

8.5.3 CUMULATIVE PROBABILITY DISTRIBUTION

For some applications, displaying the probability distribution as a cumulative function is the most useful method. A cumulative probability distribution gives the probability that a random variable x is equal to or less than a given value, a. In mathematical terms:

$$F(x) = P(x \leq a) = \int_{-\infty}^{a} f(x)\,dx$$

For example, the relative probability of T-38 miss distances from Figure 8.4 could be displayed in a cumulative distribution as in Figure 8.6.

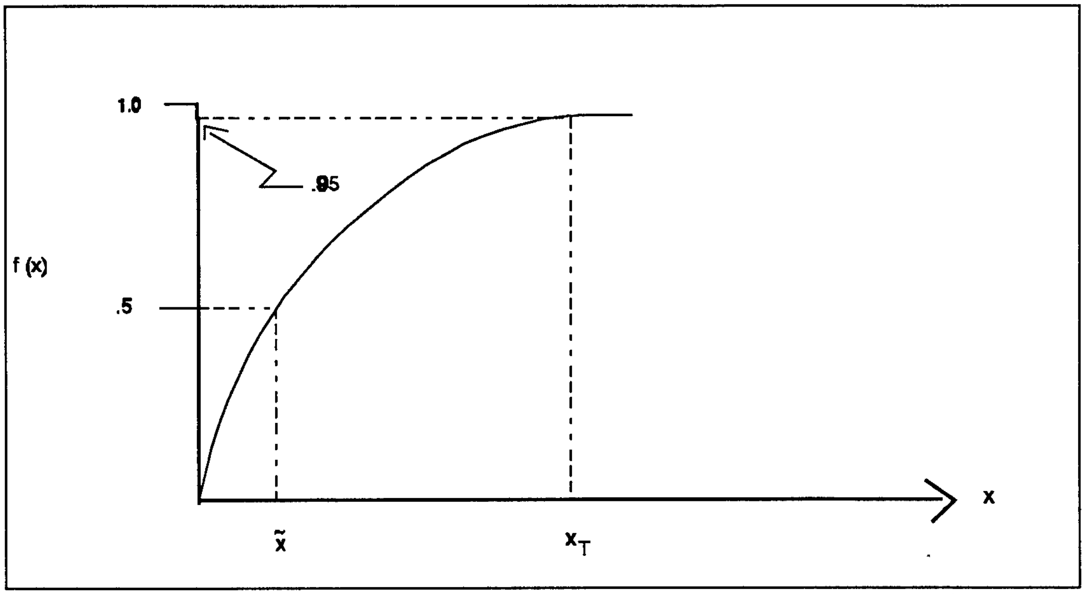


FIGURE 8.6 CUMULATIVE PROBABILITY DISTRIBUTION

From this type of display, the median, $\tilde{x}$, can be directly read. Also, we can see that 95% of the time we expect the miss distance to be below some value, x_T.

8.5.4 SPECIAL PROBABILITY DISTRIBUTIONS

There are numerous theoretically derived probability distributions used in analyzing data. In this course, we will limit our scope to only four distributions: binomial, normal, student's t, and χ^2. Each is briefly introduced below.

8.5.4.1 THE BINOMIAL DISTRIBUTION

The first special probability distribution that we will examine is a discrete probability distribution, the binomial distribution. The binomial distribution is a theoretically derived distribution of probabilities for trials in which there are two possible results, usually called success and failure. Hence the term "bi" in binomial. This really is a special case of a group of distributions called multinomial distributions. The binomial distribution can be applied to a large number of problems if success and failure are defined beforehand, for example:

1. Toss of a coin - heads (success) or tails (failure).
2. Roll of two dice - total of 7 (success) or other than 7 (failure).
3. Qualitative evaluation of a flight control modification - better (success) or worse (failure) than the original.

To determine the probability of getting exactly n successes in N trials given the probability of a single success is the problem. Let p represent the probability of a single success. First, the limiting cases are very simple.

If n = N (all successes), then the probability is just p^n. If n = 0 (all failures), then the probability is simply $(1 - p)^N$, or, if we let 1 - p = q, then q^N.

The in-between probabilities are not as simple. If we have n successes and N - n failures, we might be tempted to say that $p^n q^{N-n}$ is the probability. But there are multiple combinations of n objects possible in N events. Luckily, mathematicians have quantified how many combinations are possible and the probability of exactly n successes in N trials is:

$$f(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n}$$

where x! = x(x-1)(x-2)....(3)(2)(1) and 0! = 1. Some may recognize the first part of the right side of the above equation as the number of combinations of N things taken n at a time.

An example may help illustrate. If two different flight control systems are really equally desirable, then the probability of 6 out of 8 pilots preferring system A over system B can be found using the binomial distribution. If A and B are truly equal, the probability of a pilot picking A over B is equal (p = q = .5). The probability of exactly 6 out of 8 picking A is

$$f(6) = \frac{8!}{6!\ 2!} (.5)^6 (.5)^2 = .109$$

Thus, if you assumed that A and B were equally good, then there is only an 11% chance of getting the test results you observed, implying that your initial assumption may be in error. In a similar way, the probabilities for all possible results can be graphed as shown in Figure 8.7.

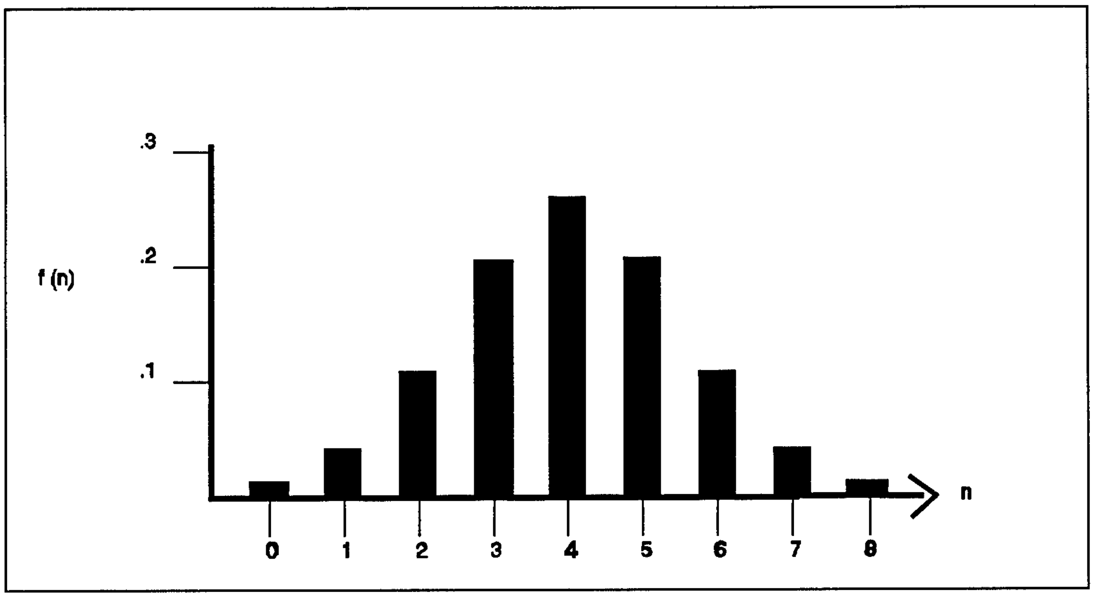


FIGURE 8.7 PROBABILITY THAT n OF 8 PILOTS WILL PREFER SYSTEM A IF P=Q=.5

To illustrate an example of the binomial distribution in which the probability of success and failure are not equal to .5, consider a production process in which the output is historically 16.7% defective. What is the probability of selecting exactly 2 defectives in a sample of 5?

p = .167

N = 5

n = 2

P(0 of 5) = .4011

P(1 of 5) = .4020

P(2 of 5) = .1612

This distribution is not symmetric like before, but is skewed to the left because of the value of p, the probability of success.

8.5.4.2 THE NORMAL DISTRIBUTION

The normal distribution is the single most important distribution in data analysis. The theoretical basis for the normal distribution lies in the binomial distribution. If we consider any deviation from the mean as the result of a large number of elemental errors, all of equal magnitude and each equally likely to be positive or negative, we can derive the following:

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Thus, the normal distribution is a continuous probability distribution, valid from $-\infty < x < \infty$. Its graphical representation is shown in Figure 8.8.

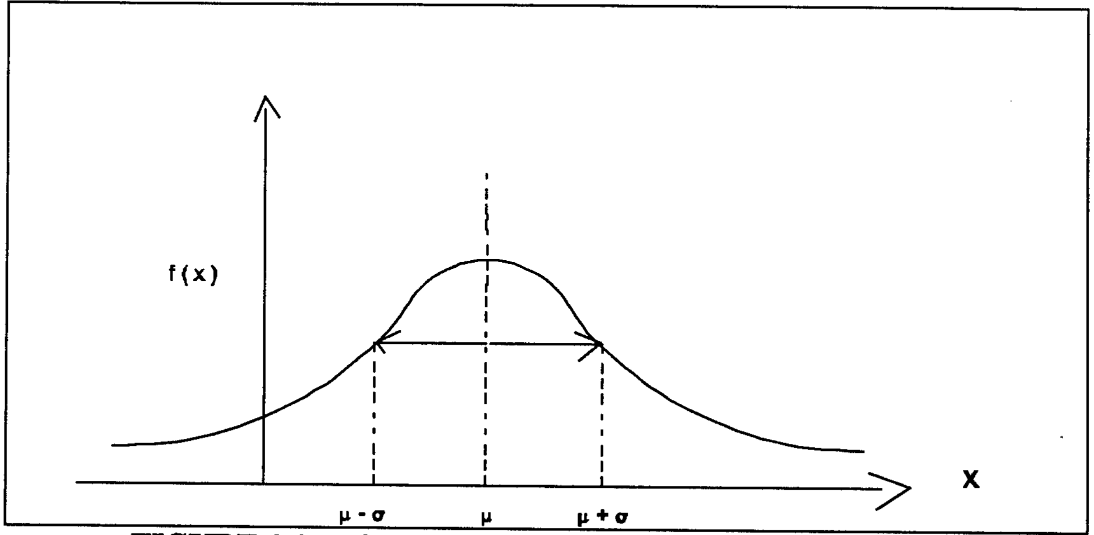


FIGURE 8.8 NORMAL PROBABILITY DISTRIBUTION

From this figure, it can be seen that f(x) is symmetric about x = μ, that x = μ yields the maximum value of f(x). Also, x = μ ± σ are the two points of inflection on the curve of f(x).

Notwithstanding the mathematical derivation of the normal distribution from a binomial distribution, the most compelling justification for its use and study is the fact that many sets of experimental observations have been shown to obey it. Accordingly, the distribution has been studied extensively.

Recalling that for a continuous probability distribution, the probability that x lies between a and b is defined by the integral of f(x) between a and b, we come to a major drawback of the normal distribution. For example, what is the probability of getting x < a if x is normally distributed? Just

$$P(x < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-(x-\mu)^2/2\sigma^2}\ dx$$

which cannot be solved in closed form. Numerical techniques are required. Tables could be used except different tables would be needed for each combination of μ and σ. The problem is overcome by making a substitution of variables in f(x) by letting

$$z = \frac{x-\mu}{\sigma}$$

then

$$f(z) = \frac{1}{\sqrt{2\pi}}\ e^{-z^2/2}$$

Tables are abundant for f(z) which is, in effect, the standardized normal distribution with a mean of zero and a standard deviation of one. To use these standardized normal tables, we must simply change our variable x to z as shown above. Values for f(z) are tabulated in Appendix D.

A graph of the standard normal distribution curve, with approximate percentages under the curve is given in Figure 8.9.

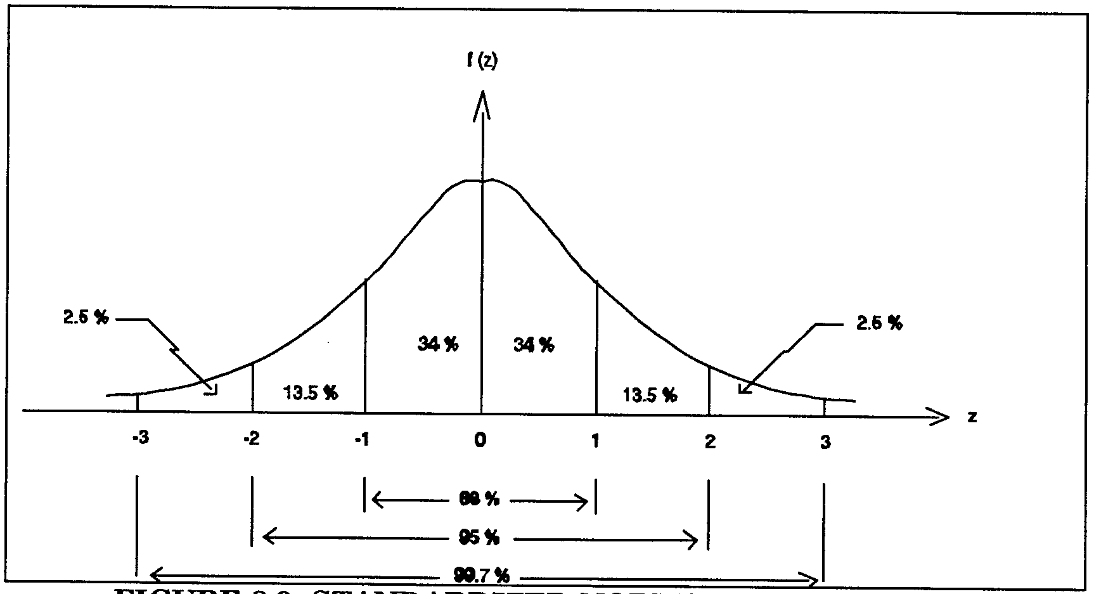


FIGURE 8.9 STANDARDIZED NORMAL DISTRIBUTION

The following three examples may help illustrate the meaning of the normal distribution and the uses of the standardized tables:

NORMAL DISTRIBUTION EXAMPLES

1. Find the area between z = .81 and z = 1.94. Using the normal distribution table in the appendix, proceed down the left column marked z until entry 1.9 is reached, then right to the column marked .04. The result, .9738, is the area between -∞ and 1.94.

Similarly, .7910 is the area from z = -∞ to .81. If we subtract this value from the first,

$$.9738 - .7910 = \underline{.1828} = P\,(.81 \leq z \leq 1.94).$$

2. Find the value of z such that the area between -1.5 and z is .0214. (Assume z is negative and to the left of -1.5.)

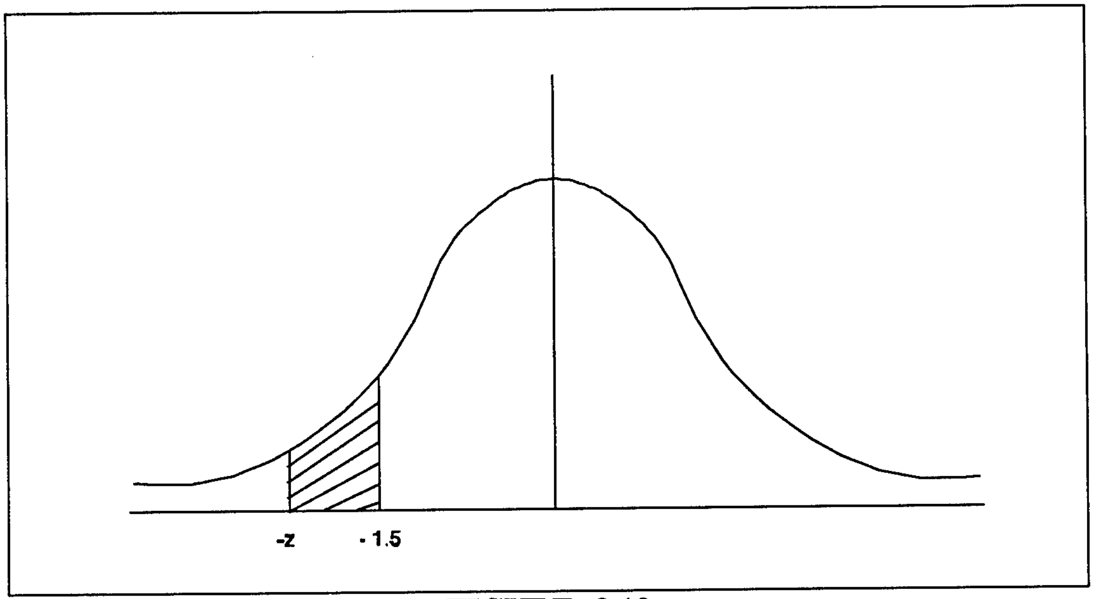


FIGURE 8.10

Area between z and -1.5 = (area between -1.5 and -∞) - (area between z and -∞)

.0214 = .0668 - (area between z and -∞)
(area between z and -∞) = .0668 - .0214 = .0454

∴ z = -1.69 (the value of z for which the area to the left of z is .0454 from the table)

3. The mean fuel used for a given profile flown 40 times was 8000 lbs, and the standard deviation was 500 lbs. Assuming the data is normally distributed, find the probability of the next sortie using between 7000 and 8200 pounds?

$$7000\ \textit{lbs in standard units} = \frac{x - \mu}{\sigma} = \frac{7000 - 8000}{500} = -2$$

$$8200\ \textit{lbs in standard units} = \frac{8200 - 8000}{500} = .4$$

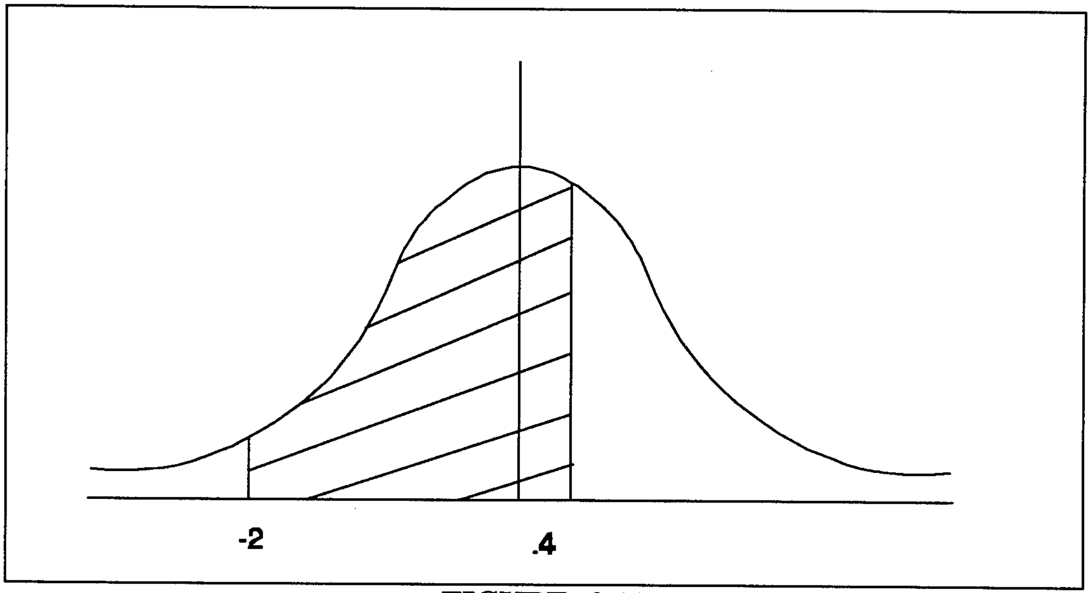


FIGURE 8.11

P(-2 ≤ z ≤ .4) = (area between z = -∞ and z = .4) - (area between z = -∞ and z = -2)
= .6554 - .0228 = .6326

8.5.4.3 THE STUDENT'S t DISTRIBUTION

In order to use the standard normal distribution, we must know the population mean and standard deviation. In practical applications, we frequently do not know these values and instead must use the sample mean and standard deviation. The difference between the sample mean and the true mean of a population was first investigated by W.S. Gossett, a brewery statistician. He developed a theoretical distribution for the statistic

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

where t is used as the measure of the difference between the sample mean and the true mean. As can be seen, the value of t is also influenced by how much dispersion we have in our sample and by the size of that sample.

For each possible value of n, we can plot a probability distribution of t. The distribution looks very similar to the standard normal distribution, especially for large values of n. In fact, it can be shown mathematically that as $n \to \infty$, the t distribution approaches the normal distribution. Figure 8.12 compares t distributions for different values of n.

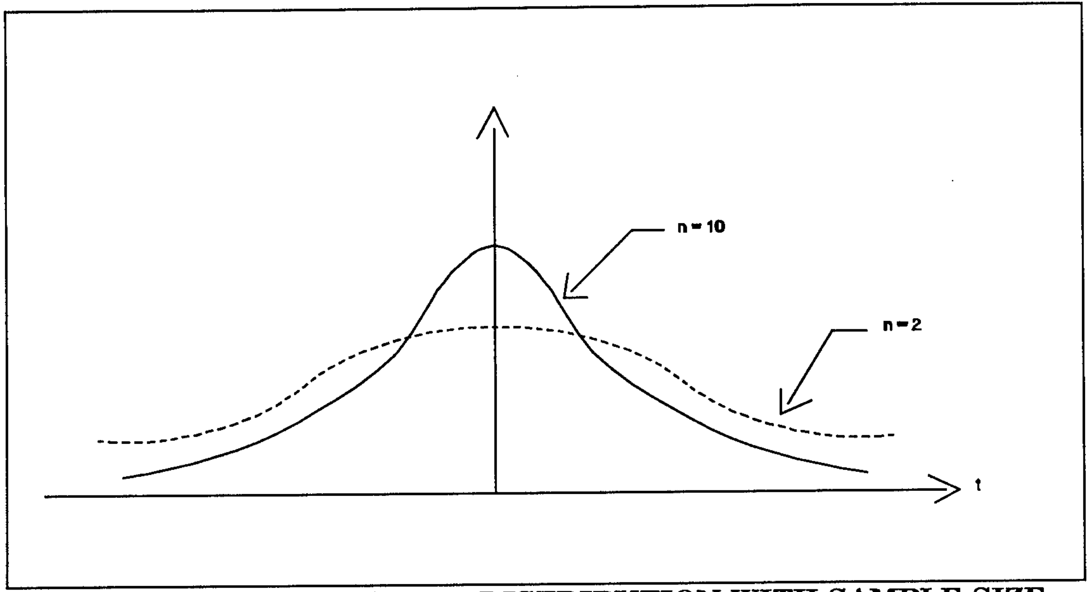


FIGURE 8.12 CHANGE IN t-DISTRIBUTION WITH SAMPLE SIZE

Because of this change in the shape of the t distribution with sample size, different t distributions must be tabulated for each value of n. Typically, as in Appendix E, different critical values of f(t) are tabulated for different values of n up to n = 30, beyond which one could use the standard normal distribution with

$$\bar{x} = \mu$$

and

$$s = \sigma$$

with very little error. It should be noted that most tables use degrees of freedom, ν, instead of n, where

$$\nu = n-1$$

The theoretical reasons for this change are of little consequence here.

8.5.4.4 THE CHI-SQUARED DISTRIBUTION

Just as the sample mean differs from the population mean, we expect the sample standard deviation to differ from the true population value. The difference is distributed according to the Chi-squared distribution of the statistic

$$\chi^2 = \frac{(n-1)\,s^2}{\sigma^2}$$

which is a measure of the dispersion of experimental s values around the population value, σ, caused by taking only limited sample sizes. A sketch of the Chi-square probability distribution is shown in Figure 8.13.

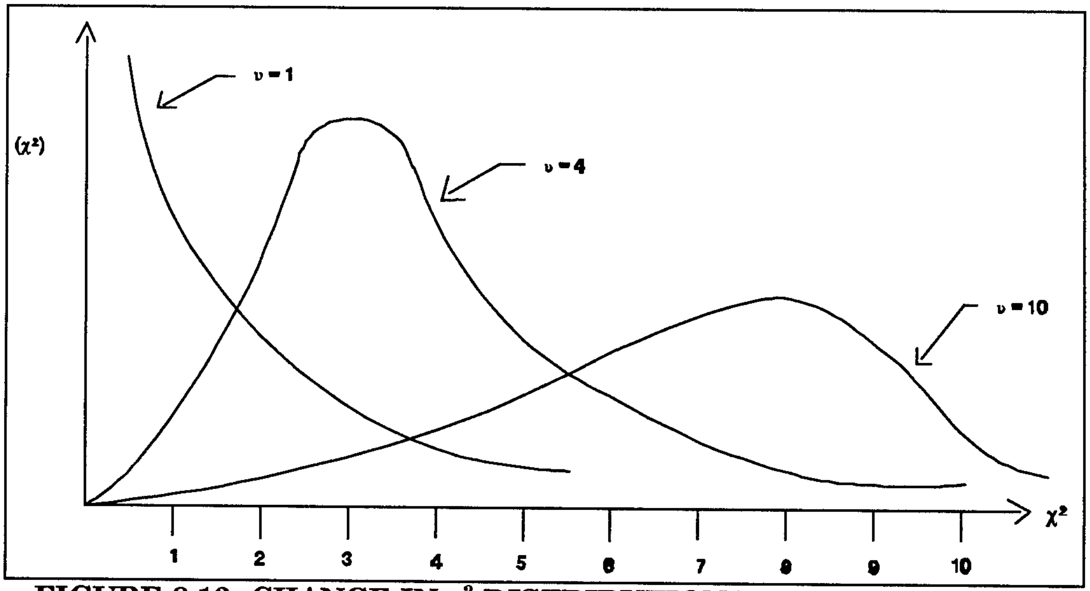


FIGURE 8.13 CHANGE IN χ^2 DISTRIBUTION WITH SAMPLE SIZE

As with the t distribution, the χ^2 distribution changes with sample size and therefore critical values of χ^2 are normally tabulated (as in the appendix) for various degrees of freedom (n - 1).

8.5.4.4.1 CHI-SQUARE DISTRIBUTION EXAMPLES

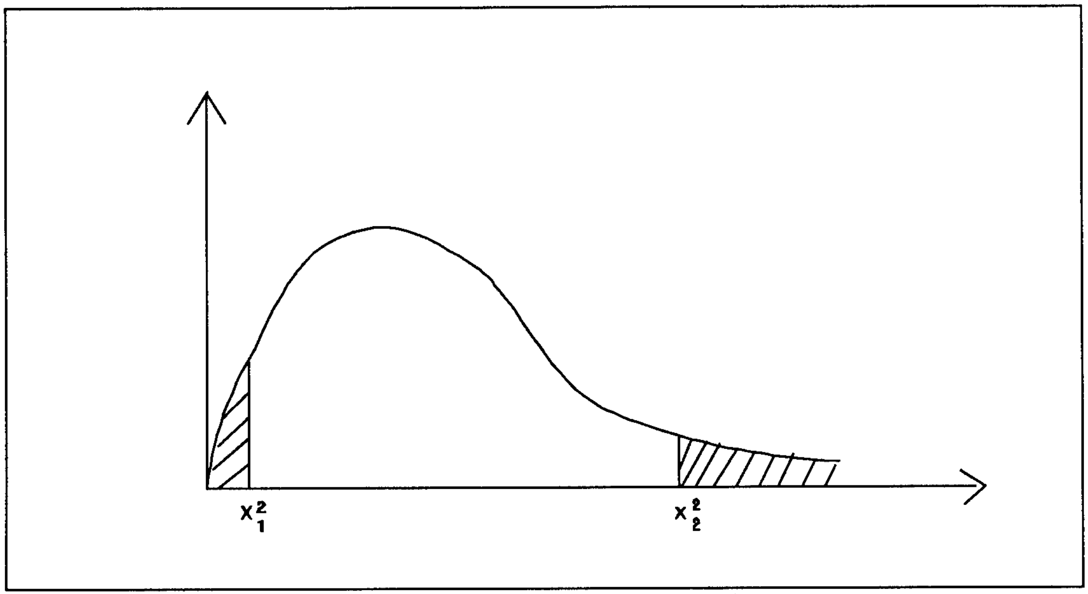


FIGURE 8.14 CHI-SQUARE EXAMPLE

1. Find the value of χ^2_2 for which the shaded area on the right = .05, assuming 5 degrees of freedom.

If the shaded area on the right is .05, then the total area to the left of χ^2_2 is (1 - .05) = .95 and χ^2_2 represents the 95th percentile, $\chi^2_{.95}$. Referring to the χ^2 distribution table in the appendix, proceed down the left column (υ) until entry 5 is reached. Then proceed right to the column headed $\chi^2_{.95}$. The result, 11.1 is the required value of χ^2_2.

2. Find χ^2_1 and χ^2_2 for which the total shaded area = .05, assuming 5 degrees of freedom.

Since the distribution is not symmetric, there are many values for which the total shaded area = .05. It is customary, unless otherwise specified, to choose the two areas to be equal. In this example, then, each area = .025.

If the shaded area on the right is .025, the area to the left of χ^2_2 is (1 - .025) = .975 and χ^2_2 is the 97.5th percentile,$\chi^2_{.975}$ which from the appendix is 12.8.

Similarly, if the shaded area on the left is .025, the area to the left of χ^2_1 is 0.025, and χ^2_1 represents the 2.5th percentile, $\chi^2_{.025}$ which equals .831.

3. Find the median value of χ^2 corresponding to 28 degrees of freedom.

Using the table in the appendix, we find in the column headed $\chi^2_{.50}$ (since the median is the 50th percentile), the value is 27.3 corresponding to $\upsilon = 28$.

8.6 CONFIDENCE LIMITS

In practical situations, we normally take a sample of a large population such as takeoff distance or bomb miss distance and we use the mean of our multiple observations as a point estimate of the true population mean. We often report this sample mean as though it were the true answer. We must realize, however, that any subsequent single observation can be expected to differ from our sample mean and that the population mean may differ from our sample mean. If we design the test correctly (standardize the method and conditions) and take sufficient samples (to be discussed in a later section), we will have confidence that our answer is sufficiently accurate. There exist statistical methods for determining how far away our answer is likely to be from the true answer (a confidence interval). These methods are the subject of this section.

8.6.1 CENTRAL LIMIT THEOREM

The central limit theorem is required to establish confidence limits on both the population mean and standard deviation. The central limit theorem can be stated as follows:

Given a population with mean, μ, and variance, σ^2, then the distribution of successive sample means, from samples of n observations, approaches a normal distribution with mean, μ , and variance σ^2/n.

In simpler terms, if we start with a general population A, where the mean is μ_A and the variance is σ^2_A, and take multiple samples, each of size n, then the resulting sample means will also have some distribution with a mean and variance ($\mu_{\bar{x}}$, $\sigma^2_{\bar{x}}$). Regardless of the original distribution of A, the distribution of the means will be approximately normal (it gets better as n is increased). Also, the mean of the means will be the same as the mean of A, and, finally, the variance of the means is the variance of A divided by $\sqrt{n}$. This is depicted in Figure 8.14.

Although proof of the central limit theorem is beyond our scope here, a cursory inspection shows that it passes the common sense test. If our sample size is very small (say 1), then for many samples, the distribution of our means is identical to the original and $\mu_{\bar{x}} = \mu_A$ and $\sigma_{\bar{x}} = \sigma_A$. At the other extreme, if n is infinite (exhaustive)

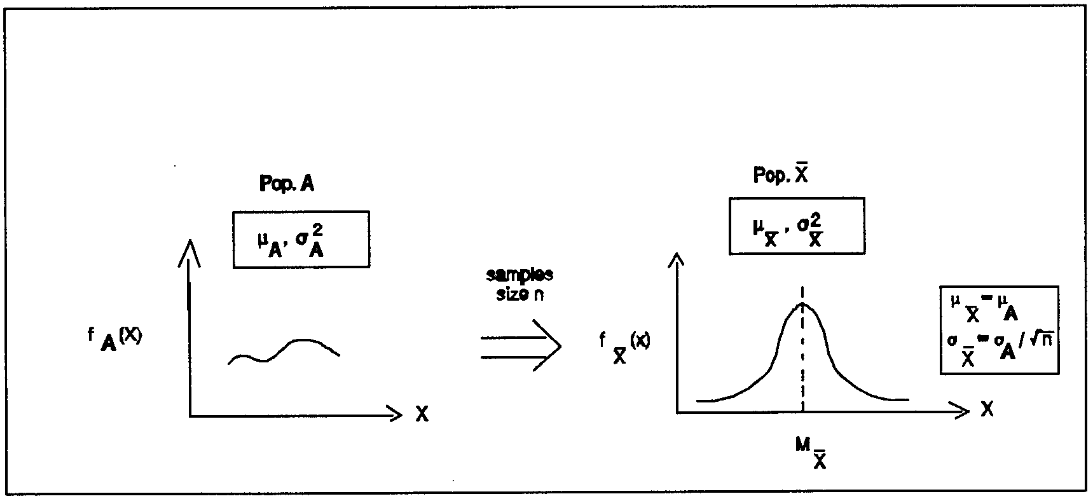


FIGURE 8.14 CENTRAL LIMIT THEOREM

then we always get the true population mean and variance. Accordingly, $\mu_{\bar{x}} = \mu_A$ $\sigma_{\bar{x}} = 0$. We now turn to using the central limit theorem to establish confidence intervals.

8.6.2 CONFIDENCE INTERVAL FOR THE MEAN

If we take a sample of size n, we now know that the distribution of the means of multiple samples would be approximately normally distributed, as shown in Figure 8.15.

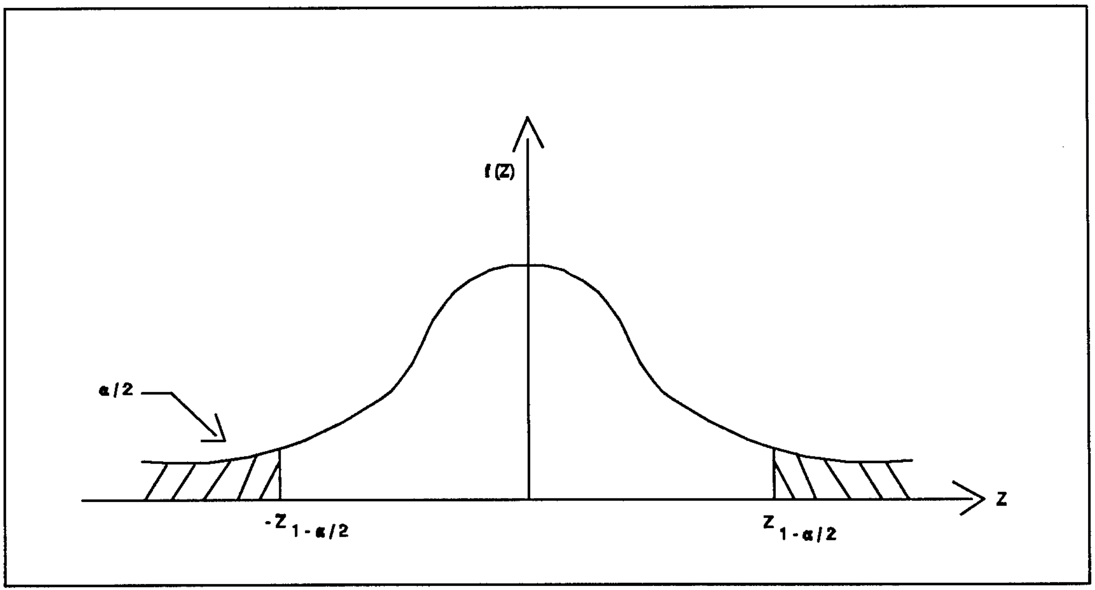


FIGURE 8.15 ESTABLISHING CONFIDENCE LIMITS ON THE MEAN

If we define α as the uncertainty level, or the percentage of time that we will be wrong, then from the definition of a normal probability distribution, we can say that a sample z will be between $-z_{1-\alpha/2}$ and $z_{1-\alpha/2}$ with probability $1 - \alpha$, or

$$P(-z_{1-\alpha/2} < z < z_{1-\alpha/2}) = 1-\alpha$$

If our z comes from one of the sample means,

$$z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

or, using the central limit theorem

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

thus

$$P\,(-z_{1-\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1-\alpha$$

or

$$P\,(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

That is, (1 - α) 100% of the time, the true population mean, μ, will be within $\pm z_{1-\alpha/2}\ \sigma/\sqrt{n}$ of the sample mean. The range of values is the interval and (1 - α) is the confidence level.

As an example, suppose we wanted to know the 95% and 99% confidence intervals for the maximum thrust of new F-100 engines given that a sample of 50 engines produced a mean max thrust of 22,700 lbs with a sample standard deviation s = 500 lbs.

1. At 95%, $\alpha = .05$ and $z_{1-\alpha/2} = 1.96$.

$$\mu = \bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Therefore,

$$\mu = 22{,}700 \pm 1.96\ (500/\sqrt{50})$$

or

$$22{,}561 < \mu < 22{,}839$$

2. At 99%, $\alpha = .01$ and $z_{1-\alpha/2} = 2.58$.

Thus,

$$\mu = 22{,}700 \pm 2.58\ (500/\sqrt{50})$$

or

$$22{,}518 < \mu < 22{,}882$$

The above example points out two important considerations. As you might have anticipated, as the requirement for certainty increases (95 → 99%), the interval widens. Given that the normal probability is continuous from $-\infty$ to $+\infty$, if we require that we be 100% certain that the true μ falls within our interval, the confidence interval becomes meaningless: $-\infty < \mu < +\infty$.

The second important point is that to construct the interval we had to use s as an estimate of σ. This is, in fact, a legitimate estimate if $n \geq 30$. For smaller sample sizes, we cannot make this assumption and must resort to the method described in the next section.

8.6.3 CONFIDENCE INTERVAL FOR MEAN FOR SMALL SAMPLES

When the sample size is less than 30 and the population variance is unknown (the typical case in flight testing), we must substitute t (defined earlier) for z:

$$(\bar{x} - t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}) < \mu < (\bar{x} + t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}})$$

As an example, suppose our earlier data on F-100 engines was based on a sample of only 5 engines. Then at the 95% confidence level:

$$\alpha/2 = .025 \text{ and } \upsilon = 4, \text{ thus } t_{4,0.975} = 2.78$$

and

$$\mu = \bar{x} \pm t_{\nu,1-\alpha/2} \frac{s}{\sqrt{n}}$$

$$= 22{,}700 \pm 2.78\ (500/\sqrt{5})$$

or

$$22{,}078 < \mu < 23{,}321$$

And as you should have expected, the interval at the same confidence level had to increase because of the smaller sample size.

8.6.4 CONFIDENCE INTERVAL FOR VARIANCE

In a manner similar to that of confidence intervals for the mean, we can establish a confidence interval for the variance based on the previously defined statistic χ^2:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

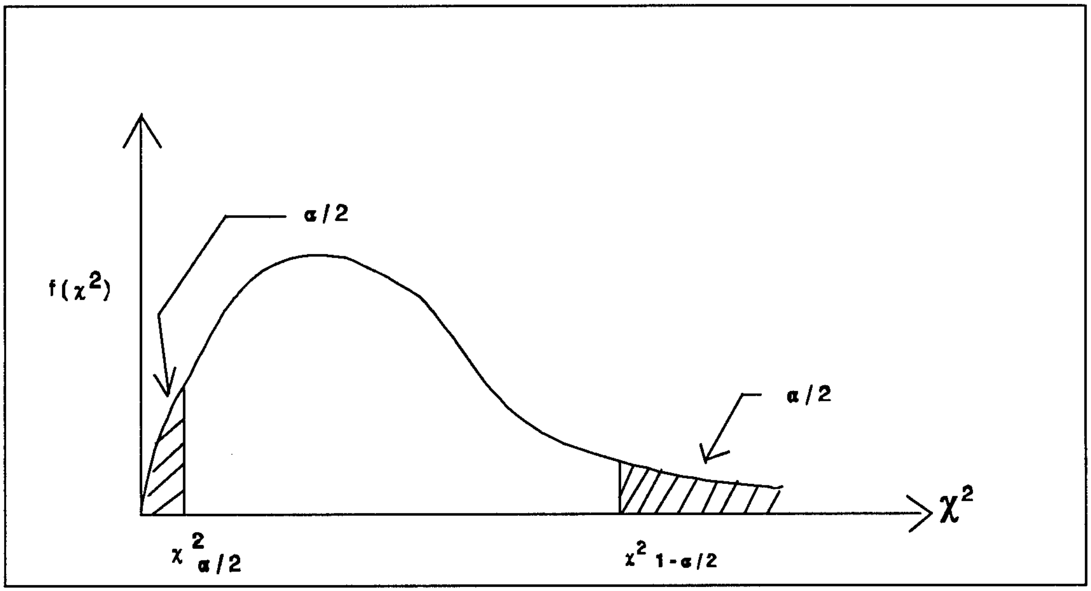


FIGURE 8.16 ESTABLISHING CONFIDENCE LIMITS ON VARIANCE

From Figure 8.16, we can see that the probability of our sample statistic, χ^2, falling between $\chi^2_{\upsilon,\alpha/2}$ and $\chi^2_{\upsilon,1-\alpha/2}$ is just 1 - α.

$$P\ (\chi^2_{v,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{v,1-\alpha/2}) = 1 - \alpha$$

Thus, with (1-α) 100% confidence,

$$\chi^2_{v,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{v,1-\alpha/2}$$

or

$$\frac{(n-1)s^2}{\chi^2_{v,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{v,\alpha/2}}$$

For example, if we take a sample of size 6 and find that the sample standard deviation is 2, we can specify with 95% probability between what limits the true population variance lies. In this case, we have:

$$\alpha/2 = .025,\ 1\text{-}\ \alpha/2 = 0.975,\ \ \upsilon = 5,\ s = 2$$

thus

$$\frac{(6-1)2^2}{\chi^2_{5,.975}} < \sigma^2 < \frac{(6-1)2^2}{\chi^2_{5,.025}}$$

where

$$\chi^2_{5,.975}=12.8,\ \chi^2_{5,.025}=.831$$

thus

$$\frac{5(4)}{12.8} < \sigma^2 < \frac{5(4)}{.831}$$

or

$$1.56 < \sigma^2 < 24.1$$

The large band is due to the small sample size. If the sample variance were the same for a larger sample (say n = 18), then the confidence interval would be smaller; for instance

$$\frac{17(4)}{30.2} < \sigma^2 < \frac{17(4)}{7.56}$$

$$2.25 < \sigma^2 < 8.99$$

8.7 HYPOTHESIS TESTING

Closely tied to the idea of confidence intervals is perhaps the most important part of statistical analysis: hypothesis testing. A statistical hypothesis is a statement, which may or may not be true, concerning one or more populations. Instead of using our sample data to make a point or interval estimate of some population parameter, we first hypothesize that a population parameter is such and such, and then use sample data to determine the reasonableness of our hypothesis. The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we examine the entire population. This is certainly the case in nearly all flight tests. A simple example may illustrate the concept.

Suppose we assume (hypothesize) that a given coin is fair, that is, the probability of heads is .5. To determine if our assumption is correct we toss the coin 100 times. If the results are 48 heads, we may conclude that it is reasonable to say the coin is fair. If, on the other hand, we get only 35 heads, it may be more reasonable to conclude that the coin is not fair. The subject of this section is how to draw the line in cases like this.

8.7.1 NULL AND ALTERNATE HYPOTHESES

It should be emphasized at the outset that the acceptance of a statistical hypothesis is a result of insufficient evidence to reject it and does not necessarily mean that it is true. Because of this fact, we must be careful in setting up our hypothesis since in the absence of data, we will be forced to accept our original hypothesis. Usually, we select this hypothesis with the sole objective of rejecting (or nullifying) it. Hence, it is called the null hypothesis, denoted H_0. The null hypothesis is usually formulated so that in the case of insufficient data, we return to the status quo or safe conclusion. Examples of null hypotheses are:

1. The defendant is innocent (not a statistical hypothesis, but a good illustration).

2. The lock-on range of a new radar is no better than that of the present radar.

3. The mean time between failures (MTBF) of a new part is no better than that of the existing part.

Since we are attempting to negate our null hypothesis, we should have established an alternate hypothesis, denoted H_1 to reflect what we want to prove and let H_0 then be the negation of H_1.

EXAMPLES:

1. H_0: $\mu = 15$ H_1: $\mu \neq 15$

2. H_0: $p \geq .9$ H_1: $p < .9$

3. H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$

8.7.2 TYPES OF ERRORS

Regardless of how carefully we set up a test, there is always the chance that we will come to the wrong conclusion. In our earlier example of tossing a coin assumed to be fair, the result of 35 heads out of 100 times could be simply due to chance variation of a fair coin (the probability of this occurring is small (.0026) but not zero). If we reject the null hypothesis when in fact it is true, this is called a Type I error, and the probability of doing so is denoted α, called the level of significance.

A different error results if we accept the null hypothesis when it is false. This is a Type II error, and its probability is denoted by β. For example, if in the coin experiment, we concluded it was a fair coin based on a result of 48 heads out of 100, the coin may really have a probability of heads of .4 and the 48 result was due to chance variation (in this case $\beta = .10$). These two different errors are summarized in Table 2. Generally, because of the fail-safe wording of the null hypothesis, we desire to have α, the probability of rejecting H_0 when it is true, very small, usually .05 (occasionally .01). The smaller α is, however, the larger β becomes. Generally, β is larger than α since this is a more acceptable error (a large β implies we stay with the status quo, H_0, more frequently than we should). The only way to reduce both α and β is to take more data. If we do exhaustive sampling, α and β go to zero.

TABLE 2. ERRORS IN HYPOTHESIS TESTING

IF	AND H_0 IS	
	True	False
Accept H_0	O.K.	Type II (β)
Reject H_0	Type I (α)	O.K.

8.7.3 ONE TAILED VS TWO TAILED TESTS

During some tests, we are interested in extreme values in either direction. Burn times on rocket motors might be an example. Too long or too short of a burn time may have dire consequences for system performance. For tests of this sort, we would form hypothesis of the form:

$$H_0: \ \mu = \mu_0 \ \text{ and } H_1: \ \mu \neq \mu_0$$

In these cases, we should reject H_0 whenever our sample produced results that were either too high or too low. Thus, our level of significance, α, would be divided into two equal regions as shown in Figure 8.17a.

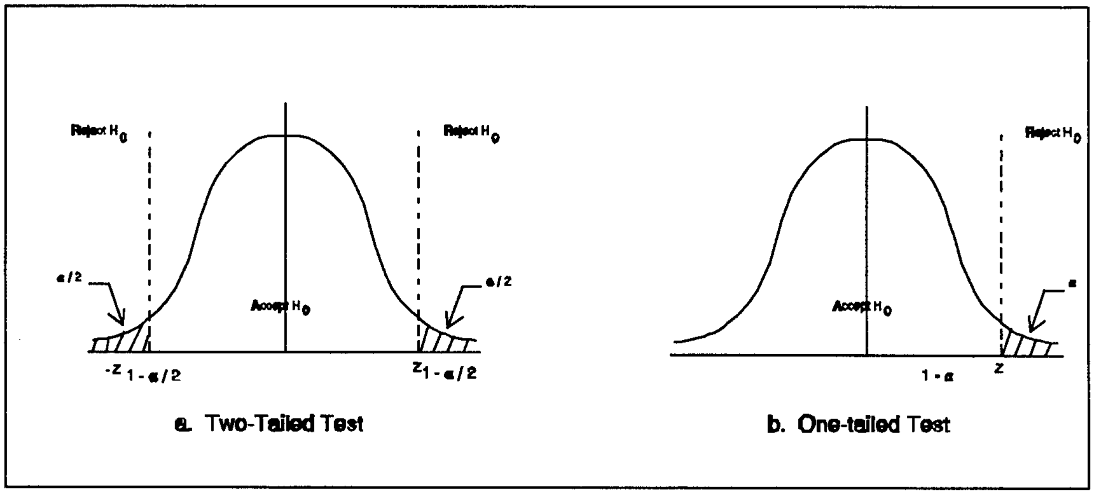
In most flight test examples, however, we are concerned with extremes in one direction only. For example, we hypothesize that the aircraft meets the contractual specification for takeoff distance. The only significant alternative hypothesis is that the actual takeoff distance is longer than the specification. For tests of this sort, we would form hypotheses of these forms:

$$H_0: \ \mu \leq \mu_0 \ \text{ and } H_1: \ \mu > \mu_0$$

or

$$H_0: \ \mu \geq \mu_0 \ \text{ and } H_1: \ \mu < \mu_0$$

In these cases, we would reject H_0 only when our sample produced results that were extreme in one direction. Thus, our level of significance, α, would be in one tail of the curve only as shown in Figure 8.17b.



a. Two-Tailed Test

b. One-tailed Test

FIGURE 8.17 ONE-TAILED VS. TWO-TAILED TESTS

8.7.4 TESTS ON MEANS

The **first step in hypothesis testing** is to formulate the null and alternate hypotheses. **Second**, choose the level of significance (α) and define the areas of acceptance and rejection. **Third**, collect data and compare the results to what was expected. **Fourth**, accept or reject the null hypothesis. For tests on means, we will use the same statistic we used in constructing confidence intervals:

1. For n > 30 or σ known, use

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

2. For n < 30 and σ unknown, use

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The following two examples should illustrate the method:

EXAMPLE 1: Two tailed test on mean, σ known. During early testing of the F-117 bombing system, it was determined that the cross range errors for 30° dive bomb passes were normally distributed with a mean error of 20 feet and a standard deviation of 3 feet. After a flight control modification to reduce adverse high AOA flying qualities, it was found that the mean cross range error for nine bomb runs was 22 feet. Has the mean changed at the .05 level of significance?

Step one: Form null and alternate hypotheses:

H_0: $\mu = 20$ (status quo)$=\rho_0$

H_1: $\mu \neq 20$

Step two: $\alpha = .05$ (given) and this will be divided into two tails, high and low, since extreme values in either direction would indicate that μ has changed.

Step three: Since s was not given, we will assume that σ has not changed significantly from the unmodified system. This is not an obvious truth, but its use here illustrates the criteria for using the z statistic. In any case, our data gives:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{22 - 20}{3 / \sqrt{9}} = 2$$

Compare this to the areas of rejection/acceptance below

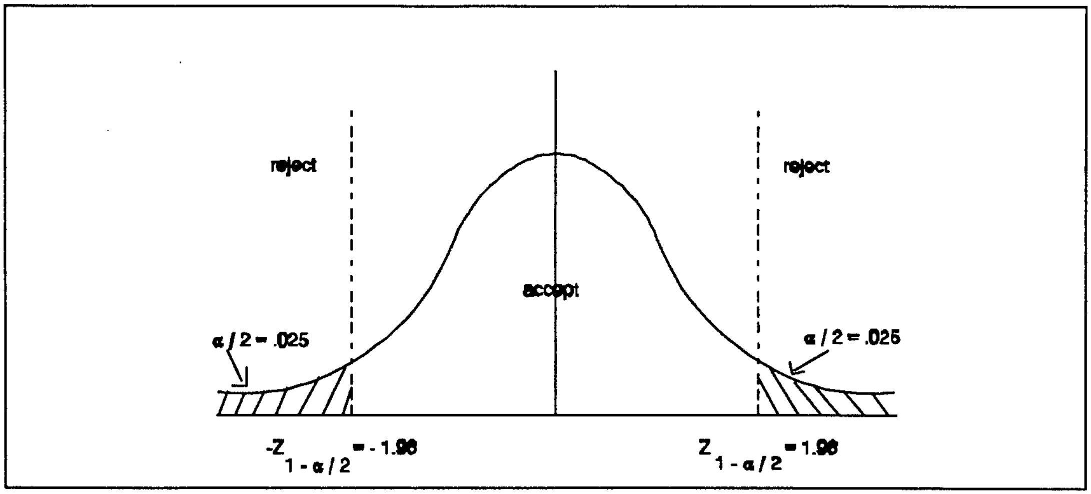


FIGURE 8.18 TWO TAILED TEST ON MEANS (σ KNOWN)

Step four: Because $z > z_{1-\alpha/2}$ $(2 > 1.96)$ we must reject the null hypothesis and conclude that (with 95% confidence) the mean cross range bombing error has changed due to the flight control modification.

EXAMPLE 2: One tailed test on mean, small sample, σ unknown. Suppose we fly nine sea level to 20,000 ft PA check climbs to verify a contract specification which states that the fuel used in this climb shall not be greater than 1500 pounds. We find that our sample of nine climbs used an average of 1600 pounds with a sample standard deviation of 200 pounds.

Step one: Form null and alternate hypotheses:

H_0: $\mu \leq 1500$ (innocent until proven guilty)

H_1: $\mu > 1500$

Step two: Choose $\alpha = .05$. An α of .01 is usually reserved for safety of flight questions. At other times, it may be specified in the contract. This is a one tailed test.

Step three: Since we have fewer than 30 data points and σ is unknown, use the data to calculate the t statistic:

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{1600-1500}{200/\sqrt{9}} = 1.5$$

Comparing this to the areas of acceptance/rejection below:

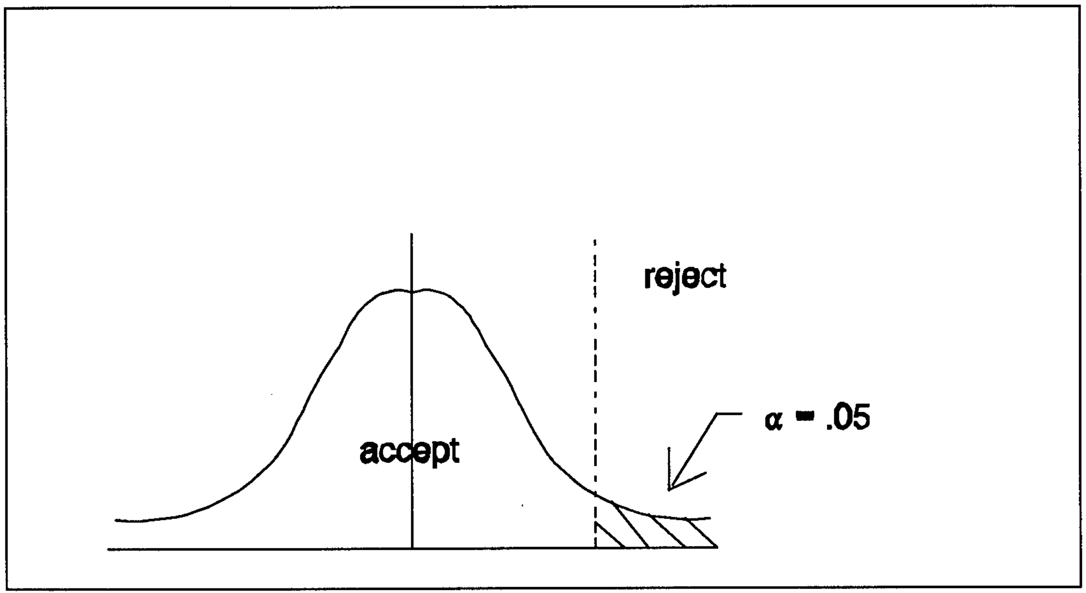


FIGURE 8.19 ONE TAILED TEST ON MEANS (σ UNKNOWN)

Step four: Because $t < t_{\upsilon,1-\alpha}$ (1.5 < 1.867) we must accept the null hypothesis and accept the contractor's claim that he has met the specification. Another way of saying it is that we don't have the data at 95% confidence to prove that the contractor has failed to meet the specification.

8.7.5 TESTS ON VARIANCE

The four steps for testing hypotheses on means described in the previous section are still valid here. The only difference in the two procedures is the use here of the chi-squared statistic instead of the z or t statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

For example, in a bombing system, the mean should be close to zero. Thus, the goodness of a system can best be measured by the dispersion of the system. Generally, the circular error probable is used as a measure of dispersion. We could, however, use the standard deviation.

Suppose the F-117 contract specification states that the standard deviation of miss distances for a particular computed delivery mode shall not exceed 10 meters at the

90% confidence level. In ten test runs, we get a standard deviation of 12 meters. Can we fine the contractor?

Step one: Form null and alternate hypotheses:

H_0: $\sigma \leq 10$

$H_{1:}$ $\sigma > 10$

Step two: An α of .10 is specified. Since smaller σ's are good, our test is a one tailed test. Only extremely large σ's will result in nullifying H_0.

Step three: Using our data, we calculate χ^2:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{9\times144}{100} = 13$$

Comparing this to $\chi^2_{v,1-\alpha}$:

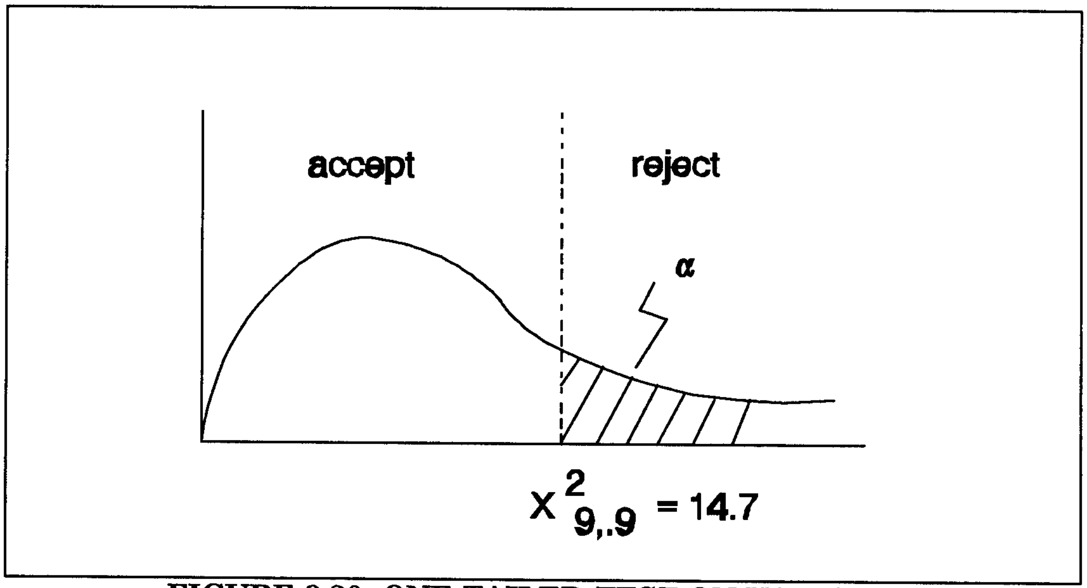


FIGURE 8.20 ONE TAILED TEST ON VARIANCE

Step four: Because $\chi^2 < \chi_{9,.9}$ (13 < 14.7) we do not have adequate data to conclude that the contractor has failed the specification. Accept H_0.

8.7.6 SUMMARY

At times, it can be a little confusing, especially with tests on variances, as to when to reject or accept the null hypotheses. Drawing figures with areas of acceptance and rejection, as has been done in the above examples, can help eliminate the uncertainty.

8.8 NONPARAMETRIC TESTS

The preceding section describes tests for populations that have normal or approximately normal distributions. Most phenomena are in fact normal. Some, however, are more accurately described by other distributions, such as the Raleigh, Cauchy, Log Normal, etc. The method of testing hypotheses described is still applicable, but the test statistic and the shape of the probability distribution would change. Tabulated values of these distributions are not always readily available. More frequently, determining the correct distribution type may be difficult. This section describes tests for populations whose distributions are not known to be normal.

8.8.1 PARAMETRIC VS NONPARAMETRIC TESTS

Nonparametric tests make no assumption concerning the shape of the population distribution. These types of tests are less powerful than the tests described in the previous section when they are used on normally distributed data. That is, they require larger sample sizes to give us the same information from the test. Because of this, the preferred procedure would be to use various tests (called goodness of fit tests) to determine the population distribution and then to use the appropriate parametric test. Failing this, a nonparametric test could be used.

Three nonparametric tests that can be useful in flight testing will be presented here: **sign test, rank sum test, and signed rank test**. The underlying basis for each of these tests is the binomial probability distribution described earlier. Essentially, each test starts out assuming that two populations are equivalent ($f_1(x) = f_2(x)$ and thus $\mu_1 = \mu_2$) and calculates statistics from two samples. Based on these test statistics, you can determine the probability of your observations, assuming identical populations. Given that probability, we can decide if our original assumption was correct.

8.8.2 THE SIGN TEST

The sign test is the simplest of the nonparametric test, and has the advantage that it can be applied to ordinal data. All that is required is paired observations of two samples with a "better than" evaluation. For example, this test can be used when each of a group of pilots evaluates two systems and identifies which he prefers.

Like the rank sum test, the null hypothesis is that the two samples came from the same population and therefore the chance of preferring System A over B is just the same as preferring B over A (i.e., .5). Therefore, here we can use the binomial distribution directly. If System A is preferred x times in N tests, the probability of this happening is:

$$f(x) = \frac{N!}{x!(N-x)!}\, p^{x} q^{N-x} = \frac{N!}{x!(N-x)!}(.5)^{N}$$

(Note that values for f(x) are tabulated in appendix 5.)

But this is a point probability in our discrete distribution, and we need the entire tail. See Figure 8.21.

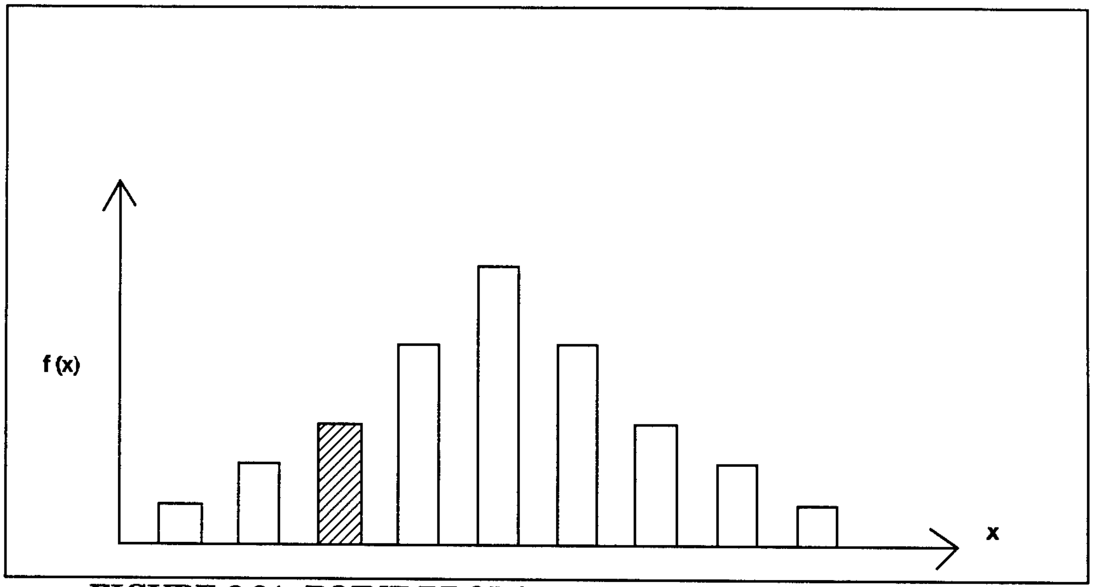


FIGURE 8.21 POINT PROBABILITY (SHADED AREA) ON A BINOMIAL DISTRIBUTION

Thus, if you need to test a single tailed hypothesis, then sum the probabilities from one end to the sample data result:

$$P(0 \leq n \leq x) = \sum_{i=1}^{x} \frac{N!}{i!(N-i)!} (0.5)^N$$

If the probability of getting a value in the tail(s) of concern is less than your chosen level of significance, then you should reject the null hypothesis that there is no difference between the systems.

For example, suppose 10 pilots evaluate the power approach handling qualities of the F-117 with two different control laws and 7 prefer System B, 2 prefer System A, and 1 has no preference. Should we switch production lines to System B? The cost is high, but if we wait to do more testing the cost will be prohibitive.

The null hypothesis is that System A and B are equally desirable. The no preference is discarded with that null hypothesis. Choose a level of significance of .05 since safety of flight is not a concern. We must now calculate the probability of getting 0, 1, or 2 pilots to chose system A if there really were no difference. If this probability is less than our level of significance, then we will reject H_0 and conclude that B is better than A.

$$P(0\ prefer\ A) = \frac{9!}{0!(9)!} (.5)^9 = .002$$

$$P(1\ prefers\ A) = \frac{9!}{1!(8)!} (.5)^9 = .018$$

$$P(2\ prefer\ A) = \frac{9!}{2!(7)!} (.5)^9 = .070$$

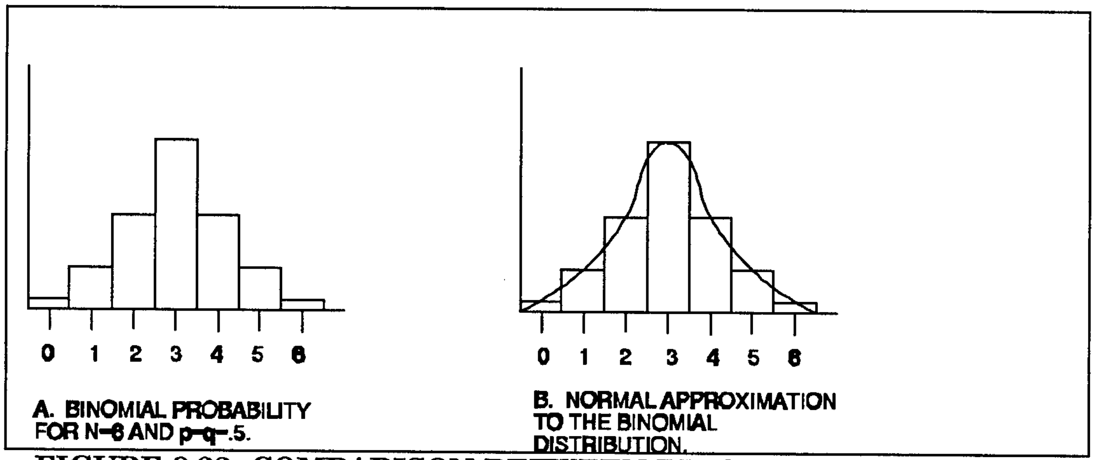
The total equals .090, or 9%.

Thus, we can only be 91% sure that B is really better than A. Not enough (at 95% significance) to justify the added expense of System B. That is, accept H_0: no significant difference between A and B.

For large sample sizes, we can use the normal approximation to the binomial distribution, provided some cautions are used. In the approximation,

$$z = \frac{x-np}{\sqrt{npq}} = \frac{x-.5n}{\sqrt{n/4}} \; for \; p = q = .5$$

Note on Figure 8.22a, the discrete binomial probability distribution for p = q = .5 and n=6 is illustrated. On Figure 8.22b, the normal approximation is overlaid on the binomial.



A. BINOMIAL PROBABILITY FOR N=6 AND p=q=.5.

B. NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION.

FIGURE 8.22 COMPARISON BETWEEN BINOMIAL AND NORMAL PROBABILITY DISTRIBUTION

From Figure 8.22b, it can be seen that when using a continuous distribution to approximate the true integer function, **the continuous function must be evaluated from the lower integer value minus 0.5 to the upper integer value plus 0.5.**

To illustrate from the previous example, with n = 9 and p = q = .5 and x = 2, we use x = 2 + 0.5 = 2.5 for the single-tail probability.

$$z = \frac{2.5-(.5)(9)}{\sqrt{9/4}}$$

From Appendix 1, this value of z yields a single-tail probability of .0918, a good approximation of the true value of .090. The accuracy improves with larger samples. When n > 15, the difference can be considered negligible.

As a further example to illustrate use of both the lower and upper integer limits, let's consider the probability of getting between 3 and 6 heads inclusive in 14 tosses of a fair coin by using the normal approximation to the binomial. The lower integer value of x (i.e., 3) minus 0.5 (= 2.5) is used for the lower limit for the normal approximation. The upper integer value of x (i.e., 6) plus 0.5 (= 6.5) is used for the upper limit. Therefore,

$$z_L = \frac{x_L-(.5)(n)}{\sqrt{n/4}} = \frac{2.5-(.5)(14)}{\sqrt{14/4}} = -2.41$$

$$z_U = \frac{x_U-(.5)(n)}{\sqrt{n/4}} = \frac{6.5-(.5)(14)}{\sqrt{14/4}} = -.27$$

$$f(z_L) = .0080$$

$$f(z_U) = .3936$$

$$f(z_U) - f(z_L) = .3936 - .0080 = .3856$$

Therefore, the normal approximation for the probability is 0.3856, which is a good approximation of the true value from Appendix 5:

$$.022+.061+.122+.183=.388$$

8.8.3 THE RANK SUM TEST

The rank sum test is also known as the U test, the Wilcoxon test, and the Mann-Whitney test in various references. This test, along with the other two nonparametric tests described in this section, can be used to test the null hypothesis that two different samples come from identical populations.

The method consists of the following steps:

1. Rank order all of the data from the two samples, noting whether each data point came from sample one or two.

2. Assign rank values to each point, one to the lowest, two to the next, etc. In the event that two or more data points have the same value, give each an average rank. For instance, if the 7th and 8th points are the same, give both a rank of 7.5.

3. Compute the sum of the ranks of each sample (R_1, R_2).

4. Calculate the following U statistics where n_1 and n_2 are sample sizes:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Note: $U_1 + U_2 = n_1 n_2$ can be used as a math check.

5. Compare the smaller U to the critical values of U listed in the appendix for $\alpha = .10$ or $\alpha = .05$.

6. If $U <$ critical value, reject H_0: $\mu_1 = \mu_2$.

While the procedure may not appear to be very intuitive, its basis is in the binomial distribution. That is, if two samples are taken from identical populations, what is the probability of getting them in a particular rank order?

As an example, consider the following. The detection ranges of two radars under controlled conditions were tested with the following results:

System 1: 9, 10, 11, 14, 15, 16, 20
System 2: 4, 5, 5, 6, 7, 8, 12, 13, 17

Is there a difference between the two systems at 95% confidence?

Using the steps described above:

1. Rank order all scores.
2. Assign rank values.

Score	4	5	5	6	7	8	9	10	11	12	13	14	15	16	17	20
System	2	2	2	2	2	2	1	1	1	2	2	1	1	1	2	1
Rank	1	2.5	2.5	4	5	6	7	8	9	10	11	12	13	14	15	16

3. Compute R_1 , R_2.

$R_1 = 7 + 8 + 9 + 12 + 13 + 14 + 16 = 79$

$R_2 = 1 + 2.5 + 2.5 + 4 + 5 + 6 + 10 + 11 + 15 = 57$

4. Calculate U_1 , U_2 .

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 7 \times 9 + \frac{7(8)}{2} - 79 = 12$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 7 \times 9 + \frac{9(10)}{2} - 57 = 51$$

5. Compare the smaller U (12 in this case) with critical values for $\alpha = .05$, $n_1 = 7$, $n_2 = 9$, $U_{cr} = 12$

6. Since U is not less than U_{cr}, we cannot reject the null hypothesis that the two radars have the same performance with 95% confidence.

8.8.4 THE SIGNED RANK TEST

The signed rank test combines elements of both the sign test and the rank sum test. Thus, the underlying assumptions are the same. System A is no better or worse than System B. Although the sign test was very simple, if we have some indication of how much better System B is than System A, then use of the sign test alone would ignore perhaps crucial data. The signed rank test incorporates this data.

The method is as follows:

1. First, rank the differences between paired observations by absolute magnitude. Ignore cases where a pair of observations is identical (i.e., no preference). Also, if there is a tie in rank order, assign an average rank to each tie.

2. Next, sum the positive and negative ranks (W_+ , W_-). The test statistic is the smaller W.

3. Compare W with critical values in the table in the appendix for the appropriate level of significance.

4. Reject H_O if $W < W_{cr}$.

As an example, suppose our previous 10 pilots who evaluated the F-117 flight control system gave systems A and B the following Cooper-Harper ratings (1 best, 10 worst):

Pilot	System A	System B	Difference
1	3	1	2
2	5	2	3
3	3	4	-1
4	4	3	1
5	3	3	0
6	4	2	2
7	4	1	3
8	2	1	1
9	3	1	2
10	1	2	-1

Ranking the differences by absolute magnitude, ignoring the zero difference gives:

Rank	2.5	2.5	2.5	2.5	6	6	6	8.5	8.5
Difference	-1	1	1	-1	2	2	2	3	3

where now

$$W_+ = 2.5 + 2.5 + 6 + 6 + 6 + 8.5 + 8.5 = 40.0$$

$$W_- = 2.5 + 2.5 = 5.0$$

Using $\alpha = .05$, W_{cr} from the tables in the appendix is 8 (using the one-tailed criteria). Since $W < W_{cr}$ ($5 < 8$), we can now reject H_0 and conclude that there is a difference between A and B with 95% confidence.

8.9 SAMPLE SIZE

All of the tests presented so far assume the data has all been collected before analysis began. Because collecting data in flight testing can be very costly in terms of money and time (there are always more things to be tested than resources allow), a scientific method to determine how many data points are needed to get statistically significant results would be very useful. We do not want our results obscured by the random variations experienced during the test. On the other hand, excessive sample sizes would give us little additional information at the expense of delaying a lower priority (but required) test.

Presented below are two approaches for determining sample size: accuracy driven and a general approach for establishing a significant difference between means.

8.9.1 ACCURACY DRIVEN

If we are required to determine a population statistic (say the mean takeoff distance) within some accuracy (say 10%), then we can use the concept of a confidence interval to determine the number of samples we need to take.

The confidence interval for the mean (σ known) is:

$$(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) \le \mu \le (\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$$

or

$$|\mu - \bar{x}| \leq z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

To achieve the select level of confidence choose:

$$n \geq \left[\frac{z_{1-\alpha/2}\ \sigma}{(\bar{x} - \mu)} \right]^2$$

but

$$|\bar{x} - \mu|$$

is the error in measuring μ. Thus, we can write

$$n \geq \left[\frac{z_{1-\alpha/2}\ \sigma}{E} \right]^2$$

For example, suppose a review of similar aircraft takeoff data shows that historically the standard deviation is about 20% of the mean. Then if the SPO wants us to determine takeoff distance to within 10% with 95% confidence, we can determine the number of times to schedule a takeoff test:

$z_{.975} = 1.96, \qquad \sigma = .2\mu, \qquad E = \pm .1\mu$

so

$$n \geq \left[\frac{(1.96)(.2\mu)}{(.1\mu)}\right]^2 = 15.4$$

Therefore, 16 sorties should be adequate to achieve the accuracy required by the SPO. As the test is in progress, we should continually check to see if our assumption concerning the standard deviation remains reasonable (tests of hypotheses on variance).

8.9.2 GENERAL APPROACH

Another frequent problem in flight testing is to determine if a system meets a specification (does $\mu = \mu_0$?) or comparing two systems to see if there is a difference (does $\mu_1 = \mu_2$?). Determining the required sample size is a lot more complex than when the criteria is simply accuracy.

Suppose we sample two different populations with means μ_1 and μ_2. As we take paired samples, we calculate the differences between them, δ. If we took a large number of samples, the resulting δ's would have some mean and distribution. If there really were no difference between the two populations, then the mean would be zero as shown in Figure 8.23. If the means were different, then the mean would be some value δ_1 as shown in Figure 8.24.

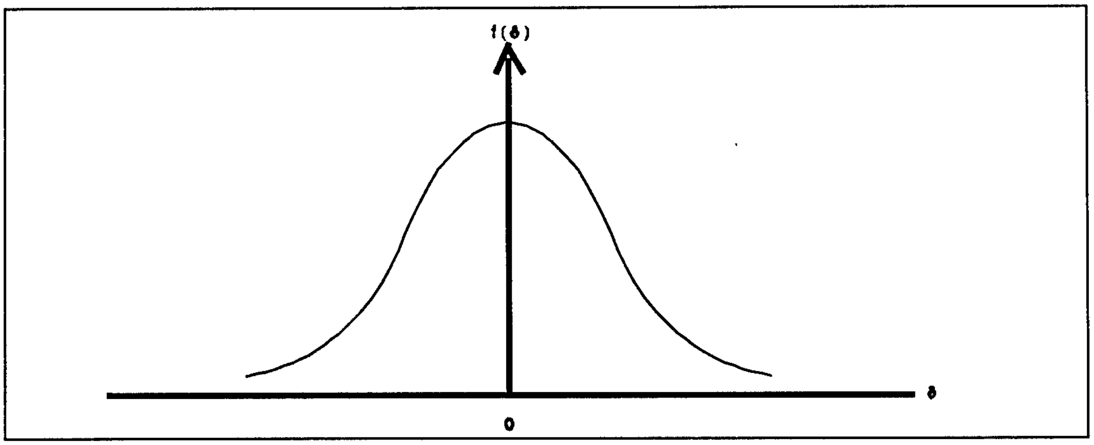


FIGURE 8.23 DISTRIBUTION OF $X_1 - X_2 = \delta$ WHEN $\mu_1 = \mu_2$

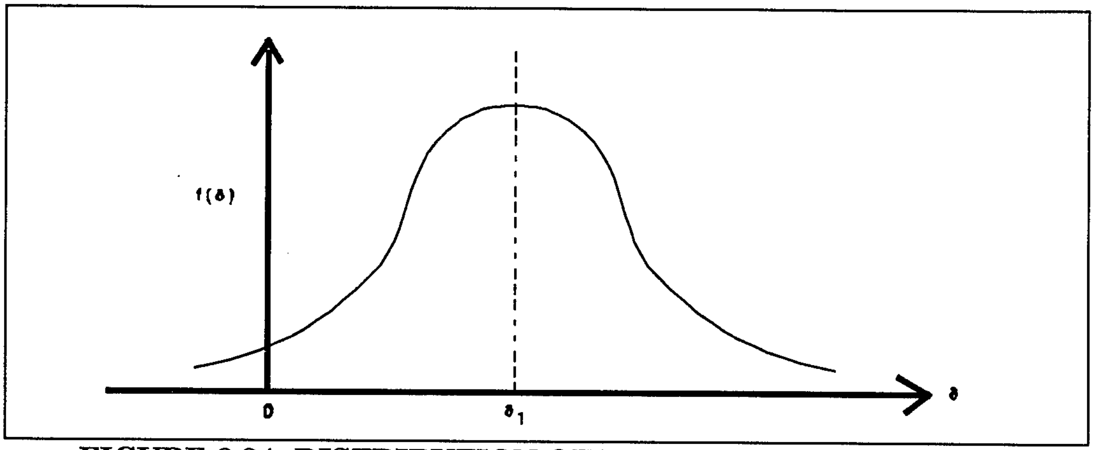


FIGURE 8.24 DISTRIBUTION OF X_1 - X_2 = δ WHEN μ_1 - μ_2 = δ_1

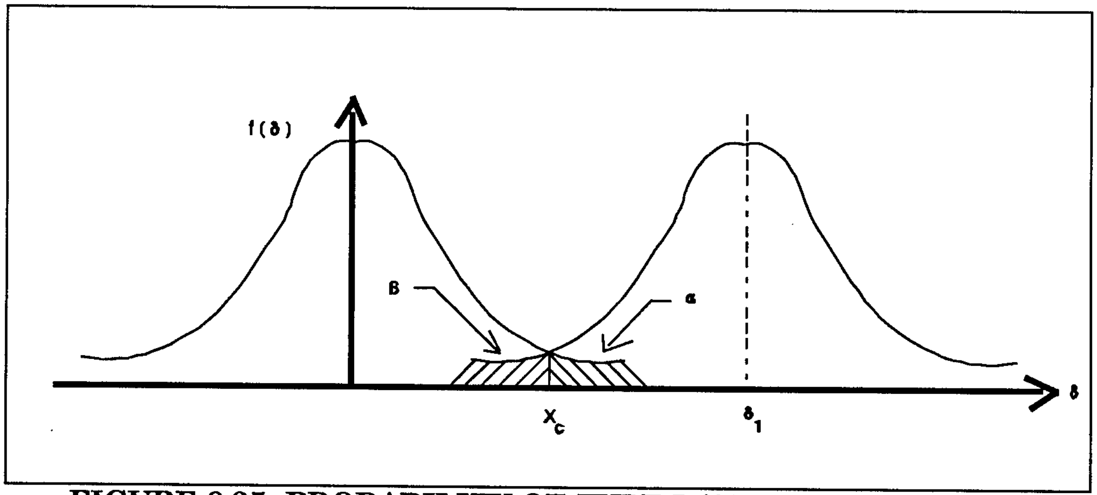


FIGURE 8.25 PROBABILITY OF TYPE I AND II ERRORS FOR COMPARING MEANS.

Combining these two alternatives in Figure 8.25, we can see that the two curves cross at some value $\delta = x_c$. A test result that gave a mean of differences above x_c would lead us to conclude that populations one and two differed in their means with level of significance of α. On the other hand, a value less than x_c would lead us to believe there was not a difference when in fact there was (with probability β as shown). The relationship between α and β can be seen graphically in Figure 8.25. If we move x_c to the right, we reduce α but increase β. Conversely, minimizing β by moving x_c left results in an increase in α. The only way to decrease α and β at the same time is to increase the sample size.

Recalling from the central limit theorem that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, we can see that α and β are direct functions of the number of samples taken and the value of δ_1. The relationship between these variables is:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2\ (\sigma^2)}{\delta_1^2}$$

The way to use this relationship is as follows:

1. Specify α. Normally, .10, .05, or .01.

2. Specify β. Usually larger than α, typically set at .10 or .20.

3. Specify δ_1. This is the least difference between μ_1 and μ_2 considered operationally significant.

4. Calculate σ. Initially, this will come from historical examples or be simply a guess. As the test continues, it can be refined.

For example, how many tests are required to determine if the contractor met the specification for a weapon delivery accuracy of 5 mils? Assume a normal error distribution with a standard deviation of 3 mils (from previous tests).

1. Set $\alpha = .05$

2. Set $\beta = .10$

3. Let $\delta_1 = 1$ mil (operationally significant)

4. $\sigma = 3$ mils

Now, we can calculate n:

$$n = \frac{(z_{.95} + z_{.90})^2 \, (3^2)}{1^2} = (1.645 + 1.28)^2 \, (3)^2 = 77$$

Thus, 77 data points are required. Practically speaking, this may be an unacceptable answer, requiring that something in 1, 2, or 3 above be changed. Tradeoffs are the subject of the next section.

8.9.3 TRADEOFFS

As can be seen from the example above, we cannot always live with our answers. In calculating n, there were many choices, some for which the consequences were not obvious. How significant is it if we change β from .10 to .20, or if we change δ_1 from 1.0 to 1.5? One good way to approach these choices is to plot the required n for various changes in α, β, and δ_1.

Then engineering judgment can be used where discretion is available. Figure 8.26 is one such example.

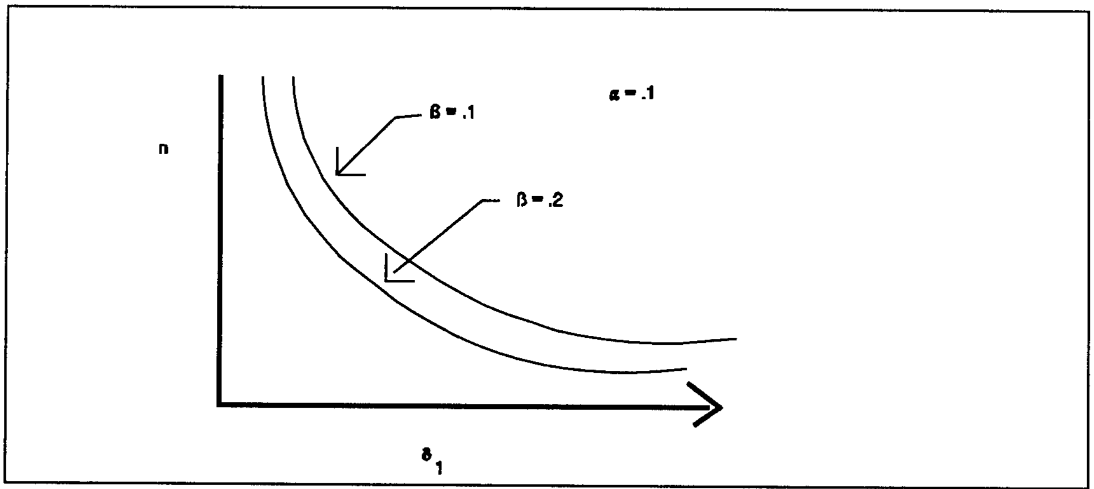


FIGURE 8.26 TYPICAL VARIATION OF SAMPLE SIZE, n, WITH MINIMUM SIGNIFICANT DIFFERENCE, δ_1

8.9.4 NONPARAMETRIC TESTS

The required sample size for nonparametric tests cannot be determined with accuracy. In practice, however, it has been found that the signed rank test is about 90% as efficient as a test on means using the z statistic. Therefore, you could calculate n as described earlier and divide by .9.

For example, how many pilots do we need to evaluate new power approach control laws in the F-117? We want to be 90% certain that there is a significant improvement (defined here as 1 Cooper-Harper rating).

1. $\alpha = .10$

2. $\beta = .20$ (arbitrary)

3. $\delta_1 = 1$

4. $\sigma = 1$

Thus,

$$n = \frac{1}{.9} \frac{(z_{.9} + z_{.8})^2 \, (1^2)}{1^2}$$

$$n = \frac{1}{.9} \, (1.28 + .84)^2 \, (1) = 4.99$$

Or 10 evaluation pilots should be planned.

8.9.5 SAMPLE SIZE MINIMIZATION

Within the principles of Design of Experiments, there are methods used to determine the minimum sample size, particularly when dealing with multi-variable tests. The normal tendency for conducting optimization tests with two variables is to hold one constant at a value and vary the other until the maximum or optimum output is achieved. Then the tendency is to set that second variable at that "optimum" value and vary the first to achieve what is hopefully the optimum output of the process.

This frequently results in erroneous conclusions. When multiple variables are involved in a test or experiment, it can be shown that there is a faster way to determine the optimum set of values for the influence factors than testing every possible mathematical combination. The Taguchi approach uses orthogonal matrices, based on the number of controllable factors in the test and the number of levels for each of these factors, to arrive at the optimum process response. These matrices represent a subset of all the possible combinations of factors and levels, but they achieve the desired results by simultaneously varying the test factors and levels.

The Taguchi approach is primarily designed to optimize product robustness (insensitivity to external influences) by experimenting with controllable combinations of influence factors. The "product" in these cases are normally the output of a stable, defined manufacturing process, such as a widget or a television. An adaption of this approach to the flight testing of an aircraft is still being developed, since the "product" of flight tests is data that show whether the aircraft or system meets its performance requirements. The product optimization method uses the following process to produce a "best" product:

1. Define the objective
2. Plan the methods of experimentation
3. Design the combination of factors and levels using an orthogonal array
4. Test the various combinations of factors and levels
5. Analyze the results of the combinations
6. Implement the changes to the process based on the results of the analysis.

The objective of applying Taguchi methods to flight testing is to obtain the required test data in the minimum number of tests. This minimization of test samples would produce the required confidence in conclusions to be drawn from the test data without expending excessive resources of time, money and material. The problem arises, however, when attempting to define the "optimum test product". For example, when testing the detection range of a radar, some of the controllable factors for the "optimum detection range" could include:

FACTOR	LEVELS	
Target radar cross section	1. large	2. small
Radar set power level	1. high	2. low
Target evasion maneuvers	1. none	2. evading
Target ECM	1. off	2. on
Background clutter	1. low	2. high
Search beam width	1. narrow	2. wide

Rather than having to test all or part of the total of all combinations of the levels and factors, the tester can intuitively know that a radar tested under all the level 1 values will have a longer detection range than one tested under level 2 conditions. It would be meaningless to conduct even a single test.

In contrast, consider a series of experiments to determine the optimum set of factor levels for growing pepper plants that will yield the hottest and largest vegetables:

FACTOR	LEVELS	
Light	1. Low	2. High
Water	1. Filtered	2. Unfiltered
Fertilizer	1. Type 1	2. Type 2
Humidity	1. Low	2. High
Type of seed	1. Brand X	2. Brand Y

In this case, the optimum combination of factor levels is not intuitive: a series of tests must be conducted to determine the combination that will yield the hottest and largest plants. The Taguchi method is used to determine the minimum number of tests, varying the levels of the controllable factors simultaneously, needed to ensure the influence of each factor and level is evaluated. The selection of the appropriate orthogonal matrix ensures that the tests will indicate the combination that maximizes plant heat and size.

8.10 ERROR ANALYSIS

Thus far in the course, we have only been concerned with the statistics of directly measured values. Often, however, measured values are used to compute some parameter of interest. For example, fuel used is usually obtained from fuel flow rate times time ($\dot{m}$ x t), and specific range is velocity divided by fuel flow ($v/\dot{m}$).

In this section, rules for determining the precision of the computed results are presented. Specifically, we will discuss significant figures, error propagation, and standard deviation of calculated values.

8.10.1 SIGNIFICANT FIGURES

The precision of an experimental result is implied by the way in which the result is written. To indicate the precision, we write a number with as many digits as are significant. The number of significant figures is defined as follows:

1. The left most nonzero digit is the most significant digit.

2. If there is no decimal point, the right most nonzero digit is the least significant digit.

3. If there is a decimal point, the right most digit is the least significant digit, even if it is zero.

4. All digits between the least and most significant digits are counted as significant digits.

For example, the following numbers each have four significant digits: 1234, 123,400; 123.4; 1001; 1000.; 10.10; 0.0001010; 100.0.

Although there are no uniform rules for deciding the exact number of digits to use when quoting measured values, the number of significant figures should be approximately one more than that dictated by the experimental precision (i.e., smallest scale division). For example, if we measure an event using a watch with tenth of a second divisions, we should not record a reading with more than two decimal places (10.24 seconds for instance). When computing a value, the following general rules apply:

1. In addition and subtraction, retain in the more accurate numbers one more decimal digit than is contained in the least accurate number (1.0 + 3.551 + 4.50 + 1.20 = 1.0 + 3.55 + 4.50 + 1.20 = 10.25).

2. In all other computations, retain from the beginning one more significant figure in the more accurate numbers than is contained in the least accurate number, then round off the final result to the same number of significant figures as are in the least accurate number (4.521/2.0 = 4.52/2.0 = 2.26 = 2.3).

When insignificant digits are dropped from a number, the last digit retained should be rounded off for the best accuracy. To round off a number to a smaller number of significant digits than are specified originally, truncate the number to the desired number of significant digits and treat the excess digits as a decimal fraction. Then

1. If the fraction is greater than 1/2, increment the least significant digit.
2. If the fraction is less than 1/2, do not increment.
3. If the fraction equals 1/2, increment the least significant digit only if it is odd.

FOR EXAMPLE: 2.53 = 2.5; 2.56 = 2.6; 2.55 = 2.6; 2.45 = 2.4

8.10.2 ERROR PROPAGATION

It should be obvious that the precision of a computed value is dependent on the precision of each directly measured value. In order to show that relationship, consider determining the volume of a right cylinder by measuring the radius and height:

$$V = \pi r^2 h$$

Given that there is some error in each measurement, call them Δr and Δh, producing some error in V, call it ΔV, then

$$V + \Delta V = \pi(r + \Delta r)^2 (h + \Delta h)$$

If the errors in r and h are small, then we can drop products of Δ's after expanding the above equation, as those products will be insignificant in comparison. This gives the following:

$$\Delta V \approx \pi(r^2 \ \Delta h + 2rh \ \Delta r)$$

or

$$\Delta V \approx \Delta h(\pi r^2) + \Delta r(2 \pi rh)$$

This grouping of the terms reminds one of partial derivatives. Specifically, it is the same as:

$$\Delta V \approx \Delta h \left(\frac{\partial V}{\partial h}\right) + \Delta r \left(\frac{\partial V}{\partial r}\right)$$

In general, it can be shown that for a function Q, where

$$Q = f(a, b, c...)$$

that the error in Q from errors in each independent variable (a, b, c...) is:

$$\Delta Q = \frac{\partial Q}{\partial a} \Delta a + \frac{\partial Q}{\partial b} \Delta b + \frac{\partial Q}{\partial c} \Delta c + \ldots$$

8.10.3 STANDARD DEVIATION

As we have seen throughout this course, we can't specify the errors in our measurements with certainty. Thus, in the place of the Δ's in the last section, a more usable equation would specify the error in the calculated parameter in terms of the standard deviation of each measured value.

From the definition of variance:

$$\sigma_Q^2 = \frac{1}{N} \sum_{i=1}^{N} (\Delta Q_i)^2$$

Using the earlier approximation for ΔQ,

$$\sigma_Q^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial Q}{\partial a} \Delta a_i + \frac{\partial Q}{\partial b} \Delta b_i + \ldots\right)^2$$

Again, dropping cross products we can write

$$\sigma_Q^2 = \frac{1}{N} \sum_{i=1}^{N} \left|\left(\frac{\partial Q}{\partial a}\right)^2 (\Delta a_i)^2 + \left(\frac{\sigma Q}{\sigma b}\right)^2 (\Delta b_i)^2 + \ldots\right|$$

Since the partial derivatives are common to each summation, they may be taken out:

$$\sigma_Q^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\partial Q}{\partial a}\right)^2(\Delta a_i)^2 + \left(\frac{\partial Q}{\partial b}\right)^2(\Delta b_i)^2 + \ldots\right]$$

$$\sigma_Q^2 = \left(\frac{\partial Q}{\partial a}\right)^2\frac{1}{N}\sum_{i=1}^{N}(\Delta a_i)^2 + \left(\frac{\partial Q}{\partial b}\right)^2\frac{1}{N}\sum_{i=1}^{N}(\Delta b_i)^2 + \ldots$$

where now the term following each partial derivation should be recognized as the definition of variance:

$$\sigma_Q^2 = \left(\frac{\partial Q}{\partial a}\right)^2\sigma_a^2 + \left(\frac{\partial Q}{\partial b}\right)^2\sigma_b^2 + \ldots$$

A complete derivation for the multivariable case can be found in Appendix B of Reference 8. As an example, consider the problem of calculating lift coefficient from the following flight test relationship:

$$C_L = \frac{841.5nW}{V_e^2S}$$

Assume that the error in S is insignificant in comparison to other errors. What is the standard deviation of CL for 1% standard deviation in each of n, W, and V_e?

First, write

$$\sigma_{C_L}^2 = \left(\frac{\partial C_L}{\partial n}\right)^2\sigma_n^2 + \left(\frac{\partial C_L}{\partial w}\right)^2\sigma_w^2 + \left(\frac{\partial C_L}{\partial V_e}\right)^2\sigma_{v_e}^2$$

or

$$\sigma_{C_L}^2 = \left(\frac{841.5W}{V_e^2S}\right)^2 (0.01n)^2 + \left(\frac{841.5n}{V_e^2S}\right)^2 (0.01W)^2 + \left(-2\frac{841.5nW}{V_e^3S}\right)^2 (0.01V_e)^2$$

or

$$\sigma_{c_L}^2 = (0.01)^2\ C_L^2 + (0.01)^2\ C_L^2 + (0.02)^2\ C_L^2$$

giving

$$\sigma_{C_L} = 0.024\ C_L$$

Thus, a 1% error in each term results in a 2.4% error in the final result.

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

1. Bethea, R. M. et. al., Statistical Methods for Engineers and Scientists, Marcel Delher, Inc., NY, 1975.

2. Young, H. D., Statistical Treatment of Experimental Data, McGraw-Hill Book Co, New York, NY, 1962.

3. Freund, J. E., Modern Elementary Statistics (6th Edition), Printice-Hall, Inc., NJ, 1984.

4. Choi, Sang C., Introductory Applied Statistics in Science, Prentice-Hall, Inc., NJ, 1978.

5. Walpole, R. E. and Myers, R. H., Probability and Statistics for Engineers and Scientists, Macmillan Co., NY, 1972.

6. Iman, R. L. and Conover, W. J., A Modern Approach to Statistics, John Wiley & Sons, Inc., NY, 1983.

7. Fletcher, D.A., "An Investigation of Taguchi Methods for Flight Test Optimization", AIAA 94-2146, 7th Biennial AIAA Flight Test Conference, Colorado Springs CO, 1984.

8. Coleman, H.W. and Steele, W.C., Experimentation and Uncertainty Analysis for Engineers, John Wiley and Sons, Inc., Canada, 1989.